Published in final edited form as:

Ann Appl Stat. 2023 June ; 17(2): 1444–1468. doi:10.1214/22-aoas1677.

CO-CLUSTERING OF SPATIALLY RESOLVED TRANSCRIPTOMIC DATA

ANDREA SOTTOSANTI[1,*], **DAVIDE RISSO**[1,†]

[1]University of Padova

Abstract

Spatial transcriptomics is a groundbreaking technology that allows the measurement of the activity of thousands of genes in a tissue sample and maps where the activity occurs. This technology has enabled the study of the spatial variation of the genes across the tissue. Comprehending gene functions and interactions in different areas of the tissue is of great scientific interest, as it might lead to a deeper understanding of several key biological mechanisms, such as cell-cell communication or tumor-microenvironment interaction. To do so, one can group cells of the same type and genes that exhibit similar expression patterns. However, adequate statistical tools that exploit the previously unavailable spatial information to more coherently group cells and genes are still lacking.

In this work, we introduce SpaRTaCo, a new statistical model that clusters the spatial expression profiles of the genes according to a partition of the tissue. This is accomplished by performing a co-clustering, i.e., inferring the latent block structure of the data and inducing two types of clustering: of the genes, using their expression across the tissue, and of the image areas, using the gene expression in the *spots* where the RNA is collected. Our proposed methodology is validated with a series of simulation experiments and its usefulness in responding to specific biological questions is illustrated with an application to a human brain tissue sample processed with the 10X-Visium protocol.

Keywords and phrases:

Co-clustering; EM algorithm; Genomics; Human dorsolateral prefrontal cortex; Integrated completed log-likelihood; Model-based clustering; Spatial transcriptomics; 10X-Visium

[*] andrea.sottosanti@unipd.it . [†] davide.risso@unipd.it .

SUPPLEMENTARY MATERIAL

Supplementary to "Co-clustering of Spatially Resolved Transcriptomic Data"
Contains the derivation of our information criterion, details on the spatial covariance functions and on the gene covariance matrices used in Section 4, details on the PCA-k-means method for selecting the number of clusters, a discussion on the computational costs of SpaRTaCo, and additional figures.

Software
Software in the form of an R package that implements SpaRTaCo is available online at https://github.com/andreasottosanti/spartaco. All the scripts to reproduce the simulations and the real data analysis are available at https://github.com/andreasottosanti/SpaRTaCo_paper.

1. Introduction.

1.1. The rise of spatial transcriptomics.

In the last few years, we have witnessed a dramatic improvement in the efficiency of DNA sequencing technologies that ultimately gave rise to new advanced protocols for single-cell RNA sequencing (scRNA-seq) and, more recently, spatial transcriptomics. In particular, spatial transcriptomics has been chosen as *method of the year 2020* (Marx, 2021). With respect to scRNA-seq, spatial transcriptomic platforms are able to provide, in addition to the abundance, the locations of thousands of genes in a tissue sample.

Righelli et al. (2021) classify spatial transcriptomic protocols into *molecule-based* and *spot-based* methods. Among molecule-based methods, seqFISH (Lubeck et al., 2014) and similar methods, such as MERFISH (Chen et al., 2015), are capable of providing the spatial expression of thousands of transcripts at a sub-cellular level, but the setup necessary to perform this kind of spatial experiments is often complex and expensive to recreate. Spot-based methods, such as Slide-seq (Rodriques et al., 2019) or the 10X Genomics *Visium* platform (Rao, Clark and Habern, 2020), have substantially lower resolution than seqFISH, but allow scientists to measure close to the whole transcriptome of (small pools of) cells across a tissue in a relatively easy manner.

Briefly, in the Visium platform, the data collection process is performed by placing a slice of the tissue of interest over a grid of spots, so that every spot contains a few neighboring cells. The gene expression of each spot is then characterized, resulting in a dataset made of tens of thousands of genes for each spot, together with the spatial location of the spots. Figure 1 shows an example of human dorsolateral prefrontal cortex (DLPFC) processed with Visium at the Lieber Institute for Brain Development (Maynard et al., 2021). The colored dots denote a manual annotation of the spots performed by Maynard et al. (2021). The dataset is available in the R package `spatialLIBD` (Pardo et al., 2021).

The rise of spatial transcriptomics has motivated the development of new statistical methods that handle the identification of *spatially expressed* (s.e.) genes, i.e., genes with spatial patterns of expression variation across the tissue. Specific inferential procedures for detecting such kind of genes, such as SpatialDE (Svensson, Teichmann and Stegle, 2018) and Trendsceek (Edsgärd, Johnsson and Sandberg, 2018), have been proposed only in the last years. These methods are widely computationally efficient, but sometimes they reach discordant inferential conclusions, and additionally they fail to account for the correlation of the genes. The very recent algorithm by Sun, Zhu and Zhou (2020), called SPARK, has addressed some of the limitations of the earlier methods. However, the additional information brought by the new spatial transcriptomic platforms has raised several questions, both on the biological and the statistical side: detecting the s.e. genes is thus not the end of the analysis but just its beginning. In this article, we want to focus on three specific research questions, i.e., to determine:

i. the clustering of the areas of the tissue sample according to the spatial variation of the genes;

ii. the existence of clusters of genes which are s.e. only in some of the areas discovered from *i.)*;

iii. the highly variable genes in the areas discovered from *i.)* net of any spatial effect.

Research question *i.)* is fundamental for the analysis of tissue samples because it is the starting point for successive downstream analyses. The recent GIOTTO (Dries et al., 2021) and BayesSpace (Zhao et al., 2021) methods are unsupervised clustering algorithms for spot-based spatial transcriptomics, designed for inferring the cell types making up a tissue. They perform a clustering based on the principle that neighboring spots are likely to be annotated with the same label, without exploiting the information carried by s.e. genes. Thus, these methods respond to a substantially different research question than *i.)*.

Research question *ii.)* is of great scientific interest, but, to the best of our knowledge, has not been tackled yet. Discovering that some genes are s.e. only in some areas of the tissue would play a core role in comprehending some fundamental biological mechanisms, and ultimately discovering new ones. Even the very recent SPARK method for detecting s.e. genes is not designed to state if the spatial expression activity of a gene is restricted to specific areas of the tissue. With the existing statistical tools, one can approach this issue with a two-step analysis, first clustering the image using BayesSpace or GIOTTO, and then applying SPARK to each of the discovered clusters. However, such heuristic procedure has some severe limitations. First, repeating the tests in each of the image cluster requires to control for multiple testing, e.g., by controlling the False Discovery Rate (Benjamini and Hochberg, 1995). Second, even after the s.e. genes are isolated, an additional clustering of the genes is necessary to perform specific downstream analyses (Svensson, Teichmann and Stegle, 2018; Sun, Zhu and Zhou, 2020). Last, if indeed there are clusters of genes, such information should be accounted for in the first step of the procedure, when the image is clustered. However, this is something that cannot be accomplished with BayesSpace or GIOTTO.

Finally, research question *iii.)* has the goal of determining which genes are active in each of the image cluster. Thanks to the spatial mapping of the spots, it will be possible to separate the presence of spatial effects from the total variation of each gene, providing a more accurate list of highly variable genes.

1.2. A co-clustering perspective.

In this article, we consider the problem of modelling and clustering gene expression profiles in a tissue sample processed with a spot-based spatial transcriptomic method, such as 10X Visium, and measured over a set of spatially located sites.

In the remainder of the article, we use “spots” to denote the spots in the tissue from which RNA is extracted and “genes” to denote the variables measured in each spot, using a terminology typical of the Visium platform. However, the method presented here is more general and can be applied to any spatial transcriptomic technology and, more broadly, to any dataset for which the rows or the columns are measured in some observational sites with known coordinates.

We tackle the research questions outlined above as a single, two-directional clustering problem: of the genes, using spots as variables, and of the spots, using genes as variables. This kind of procedure is known in the literature as *co-clustering* (or *block-clustering*, Bouveyron et al., 2019) and denotes the act of clustering both the rows and the columns of a data matrix, which, in this way, is partitioned into rectangular, non-overlapping sub-matrices called *co-clusters* (or *blocks*).

Bouveyron et al. (2019) distinguish between *deterministic* and *model-based* co-clustering approaches. Model-based methods are designed to simultaneously perform the clustering and reconstruct the probabilistic generative mechanism of the data. The model-based co-clustering literature is centered around the Latent Block Model (LBM; Govaert and Nadif, 2013), an extension of the standard mixture modelling approach when both rows and columns of a data matrix are deemed to come from some underlying clusters. Thanks to the ease of interpretation and to the raise of new advanced computational methods, the LBM has been extensively explored as a tool for modelling continuous (Govaert and Nadif, 2013, Chapter 5), categorical (Keribin et al., 2015), count (Govaert and Nadif, 2010), binary (Govaert and Nadif, 2008) and recently even functional data (Bouveyron et al., 2018; Casa et al., 2021). In addition, both frequentist (Govaert and Nadif, 2008; Bouveyron et al., 2018) and Bayesian (Wyse and Friel, 2012; Keribin et al., 2015) approaches have been proposed for fitting these models. The conditional independence assumption of LBM states that the observations within the same co-cluster are independent. Surely, this hypothesis is computationally attractive, yet it is incompatible with the high correlation levels shown by gene expression data (Efron, 2009).

Tan and Witten (2014) overcome the conditional independence assumption proposing a co-clustering model based on the matrix variate Gaussian distribution (Gupta and Nagar, 2018), which accounts for the dependency across the rows and the columns in a block with two non-diagonal covariance matrices. Their model represents a first attempt to extend k-means-type algorithms for co-clustering to the case where the data entries in a block are not independent. The estimation of the needed covariance matrices is challenging; a challenge that can be overcome with the aid of a penalization term, such as the LASSO (Witten and Tibshirani, 2009), to avoid singularity problems. However, with spatial data, it is natural to leverage the spatial dependencies observed in the data to aid the covariance matrix estimation.

Here, we propose SpaRTaCo (SPAtially Resolved TrAnscriptomics CO-clustering), a novel co-clustering technique designed for discovering the hidden block structure of spatial transcriptomic data. Since the spots in which gene expression is measured are spatially located on a grid, our model expresses the correlation across transcripts in different spots as a function of their distances. As a consequence, differently from the rest of the co-clustering models proposed in the literature, SpaRTaCo divides the data matrix into blocks based on the estimated means, variances, and spatial covariances. In addition, we use gene-specific random effects to account for the remaining covariance not explained by the spatial structure.

Although the published literature is not always clear about the distinction between *co-clustering* and *biclustering*, in accordance with the recent works of Moran, Ročková and George (2021) and Murua and Quintana (2021) here we adopt the following terminology: both co-clustering and biclustering are families of techniques used to group the rows and the columns of a data matrix. However, in biclustering the groups formed, called *biclusters*, can take any possible shape, while co-clustering is limited to rectangular, non-overlapping blocks. In addition, biclustering algorithms do not necessarily allocate all the data entries into one of the existent biclusters, and so some entries can be left unassigned. Although biclustering methods are more flexible, the main advantage of co-clustering is that the returned blocks are often easier to interpret both from a statistical and practical perspective.

1.3. Outline.

The rest of the manuscript is structured as follows. Section 2 illustrates the SpaRTaCo modelling approach and reviews some competing co-clustering models, highlighting the similarities and the differences with our proposal. Section 3 discusses some identifiability issues, illustrates our classification-stochastic EM (CS-EM) algorithm for parameter estimation, proposes a measure to quantify the clustering uncertainty, and derives a model selection criterion based on the *integrated completed log-likelihood* (Biernacki, Celeux and Govaert, 2000). Section 4 proposes five simulated spatial experiments of growing complexity with whom we compare SpaRTaCo with other co-clustering models. Section 5 shows how our proposal allows to answer our three research questions using the human brain tissue sample displayed in Figure 1. The manuscript is concluded by some considerations of the possible future extensions.

2. The statistical model.

Let $\mathrm{X} = (x_{ij})_{1 \le i \le n, 1 \le j \le p}$ be the $n \times p$ matrix of a spatial experiment processed by a spot-based spatial transcriptomic platform, i.e, containing the expression of n genes over a grid of p spots on the chip surface. The spatial location of the spot j over the chip surface is known through its spatial coordinates $\mathrm{s}_j = (s_{jx}, s_{jy})$; we name as $S = (s_j)_{1 \le j \le p}$ the $p \times 2$ matrix containing the coordinates of the p spots. From this point, we assume that the data entries in X have been properly pre-processed, and so $x_{ij} \in \mathbb{R}$ for any i and j (see Section 5).

2.1. Model formulation.

We assume there exist K clusters of rows of X, and R clusters of columns of X, forming a latent structure of KR blocks. The vectors of random variables $\mathscr{Z} = (\mathscr{Z}_i)_{1 \le i \le n}$ and $\mathscr{W} = (\mathscr{W}_j)_{1 \le j \le p}$ denote to which cluster the rows and the columns belong, respectively. Thus, $\mathscr{C}_k = \{i = 1, \ldots, n : \mathscr{Z}_i = k\}$ is the k-th row cluster, with $k = 1, \ldots, K$, and $\mathscr{D}_r = \{j = 1, \ldots, p : \mathscr{W}_j = r\}$ is the r-th column cluster, with $r = 1, \ldots, R$. The cluster dimensions are $n_k = |\mathscr{C}_k|$ and $p_r = |\mathscr{D}_r|$. The notation used to refer to subsets of X is the following: $\mathrm{X}^{kr} = (x_{ij})_{i \in \mathscr{C}_k, j \in \mathscr{D}_r}$ is the kr-th co-cluster (block), $\mathrm{X}^{k\cdot} = (x_{ij})_{i \in \mathscr{C}_k, 1 \le j \le p}$ is the $n_k \times p$ matrix formed by all the rows in $\mathscr{C}_k$, and $\mathrm{X}^{\cdot r} = (x_{ij})_{1 \le i \le n, j \in \mathscr{D}_r}$ is the $n \times p_r$ matrix formed by all the columns in $\mathscr{D}_r$. When it comes to access the elements of a block, we use the notation

$X^{kr} = \left(x_{ij}^{kr}\right)_{1 \le i \le n_k, 1 \le j \le p_r}$. So, the i-th row vector and the j-th column vector of X^{kr} are respectively $x_{i.}^{kr} = \left(x_{ij}^{kr}\right)_{1 \le j \le p_r}$ and $x_{.j}^{kr} = \left(x_{ij}^{kr}\right)_{1 \le i \le n_k}$.

The vector $x_{i.}^{kr}$ contains the expression of the i-th gene in the cluster $\mathscr{C}_k$ across the p_r spots in the cluster $\mathscr{D}_r$. We model $x_{i.}^{kr}$ as

$$x_{i.}^{kr} = \mu_{kr} 1_{p_r} + \sigma_{kr,i}\epsilon_{i.}^{kr}, \quad \epsilon_{i.}^{kr} \sim \mathcal{N}_{p_r}(0, \Delta_{kr}), \tag{1}$$

$$\Delta_{kr} = \tau_{kr}\mathscr{K}\left(S^r; \phi_r\right) + \xi_{kr}\mathbb{I}_{p_r}, \tag{2}$$

where μ_{kr} is a scalar mean parameter, 1_{p_r} is a vector of ones, $\sigma_{kr,i}^2$ is a gene-specific variance, and Δ_{kr} is the covariance matrix of the columns. Following Svensson, Teichmann and Stegle (2018) and Sun, Zhu and Zhou (2020), Formula (2) expresses Δ_{kr} as a linear combination of two matrix terms: $\mathbb{I}_{p_r}$ is a diagonal matrix of order p_r, $\mathscr{K}\left(S^r;\phi_r\right) = \left(k(\|s_j^r - s_{j'}^r\|;\phi_r)\right)_{1 \le j, j' \le p_r}$ is the spatial covariance matrix, where $k(\cdot;\phi_r)$ is an *isotropic* spatial covariance function (Cressie, 2015) parametrized by a vector ϕ_r, and $S^r = (s_j)_{j \in \mathscr{D}_r}$ is the sub-matrix of S containing the spots in $\mathscr{D}_r$. The term isotropic denotes that the covariance between two points $j, j' \in \mathscr{D}_r$ depends just on the distance between their two sites, $\|s_j^r - s_{j'}^r\|$. The positive parameters τ_{kr} and ξ_{kr} in Formula (2) handle the linear combination between $\mathscr{K}$ and $\mathbb{I}_{p_r}$: the former measures the spatial dependence of the data, the latter is the so-called *nugget effect*, a residual variance.

According to Section 2.4 of Cressie (2015), to select an adequate spatial covariance kernel for the data, one can explore the empirical spatial dependency through the *variogram* and then select a kernel from a vast list of proposals (see for example Rasmussen and Williams, 2006). However, under our model, this strategy would be unfeasible because only the columns within the same cluster are spatially dependent, so the selection of the spatial covariance kernel should be performed simultaneously with the clustering of the data. As a compromise, SpaRTaCo considers the same covariance model $k(\cdot;\phi_r)$ for every column cluster $\mathscr{D}_r$; the only difference among the kernels of the clusters is the value of the model parameters ϕ_r.

The scale parameters $\sigma_{kr,i}^2$ in (1) aim to capture the variability left unexplained by the spatial covariance model (2), and possibly the extra source of variability due to the dependency across genes. In the longitudinal data framework, De la Cruz-Mesía and Marshall (2006) and Anderlucci and Viroli (2015) consider a random effect model to account for the systematic dependency across subjects in the same group of study. We follow the same approach and we assume that every $\sigma_{kr,i}^2$ is a realization of an Inverse Gamma distribution $\mathscr{IG}(\alpha_{kr}, \beta_{kr})$, where α_{kr} and β_{kr} denote the shape and the rate, respectively. The Inverse Gamma is chosen for its conjugacy with the Gaussian distribution and allows to derive the marginal probability density of $x_{i.}^{kr}$, that is

$$f\left(\mathrm{x}_{i.}^{kr};\theta_{kr},\phi_r\right)=\frac{1}{\sqrt{(2\pi)^{p_r}\det(\Delta_{kr})}}\frac{\Gamma(\alpha_{kr,i}^{*})}{\Gamma(\alpha_{kr})}\frac{\beta_{kr}^{\alpha_{kr}}}{\beta_{kr,i}^{*}\alpha_{kr,i}^{*}}, \tag{3}$$

where $\det(\cdot)$ denotes the matrix determinant, $\alpha_{kr,i}^{*}=p_r/2+\alpha_{kr}$ and $\beta_{kr,i}^{*}=\left(x_{i.}^{kr}-\mu_{kr}1_{p_r}\right)^{T}\Delta_{kr}^{-1}\left(x_{i.}^{kr}-\mu_{kr}1_{p_r}\right)/2+\beta_{kr}$. Note that this formulation corresponds to the probabilistic model $x_{i.}^{kr}\sim t_{2\alpha_{kr}}\left(\mu_{kr}1_{p_r},\alpha_{kr}^{-1}\beta_{kr}\Delta_{kr}\right)$ and is similar to that employed to *shrink* the gene variances in the popular *limma* model (Smyth, 2004). The set of parameters $\theta_{kr}=\{\mu_{kr},\tau_{kr},\xi_{kr},\alpha_{kr},\beta_{kr}\}$ is specific of the data into the (k,r)-th co-cluster, while ϕ_r is a parameter that is descriptive of the entire r-th column cluster.

The model in Formula (1) can be rephrased with a probability distribution over the entire kr-th block, $\mathrm{X}^{kr}|\Sigma_{kr}\sim\mathscr{MVN}(\mu_{kr}1_{n_k\times p_r},\Sigma_{kr},\Delta_{kr})$, where $\mathscr{MVN}$ denotes the matrix-variate normal distribution and $\Sigma_{kr}=\mathrm{diag}\left(\sigma_{kr,1}^{2},\ldots,\sigma_{kr,n_k}^{2}\right)$ is the (diagonal) covariance matrix of the genes. A consequence of the matrix-variate normal model is that every row, column and sub-matrix of X^{kr} is Gaussian (Gupta and Nagar, 2018). For instance, the following model formulation is equivalent to Formula (1):

$$x_{.j}^{kr}|\Sigma_{kr}\sim\mathscr{N}_{n_k}\left\{\mu_{kr}1_{n_k},(\tau_{kr}+\xi_{kr})\Sigma_{kr}\right\},\quad \mathrm{Cov}\left(x_{.j}^{kr},x_{.j'}^{kr}\right)=\tau_{kr}k(\|s_j^r-s_{j'}^r\|;\phi_r)\Sigma_{kr},$$

with $j,j'\in\mathscr{D}_r$.

Last, the clustering labels $\mathscr{Z}$ and $\mathscr{W}$ are unknown independent random variables. Figure 2 represents the relations across the elements of the model with a DAG.

2.2. A comparison with other co-clustering models.

We review in this section some advanced co-clustering techniques that have some similarities with our proposal. The goal is to highlight, starting from the existing literature, how SpaRTaCo has been designed specifically for detecting and clustering data based on their spatial covariance in some groups of observational sites. With respect to the distinction between deterministic and model-based co-clustering techniques we already discussed in Section 1.2, we choose to compare SpaRTaCo only with model-based techniques because they offer a clear advantage in the interpretation of the results. Some of the methods that we review here are named as biclustering models, but in practice they segment the data matrix into rectangular blocks.

Sparse Biclustering (sparseBC, Tan and Witten, 2014) extends the *k-means* algorithm to the co-clustering framework. The model corresponds to a probabilistic assumption on the block of the type $\mathrm{X}^{kr}\sim\mathscr{MVN}(\mu_{kr}1_{n_k\times p_r},\mathbb{I}_{n_k},\xi\mathbb{I}_{p_r})$, where ξ is an unknown scale parameter. In sparseBC, the estimation of μ_{kr}, for any k and r, is regulated by a LASSO penalization. We thus distinguish the sparse estimation from the case of null penalization (BC).

Matrix-Variate Normal Biclustering (MVNb, Tan and Witten, 2014) extends sparseBC by taking a probabilistic model on the blocks of the type $\mathrm{X}^{kr}\sim\mathscr{MVN}\left(\mu_{kr}1_{n_k\times p_r},\Sigma_k^{\mathrm{MVNB}},\Delta_r^{\mathrm{MVNB}}\right)$, where both Σ_k^{MVNB} and Δ_r^{MVNB} are non-diagonal covariance matrices with respectively $n_k(n_k + 1)/2$ and $p_r(p_r + 1)/2$ free parameters. Together with the LASSO penalization on the centroids, handled by a parameter λ, the authors deploy a graphical LASSO penalization (Witten and Tibshirani, 2009) to practically solve the singularity problems in the estimate of Σ_k^{MVNB} and Δ_r^{MVNB}. The penalization parameters involved are denoted by ρ_Σ and ρ_Δ. With respect to the MVNB, SPARTACO has specific row and column covariance matrices Σ_{kr} and Δ_{kr} for each block, whose structure is described in Section 2.1. The total number of free parameter, $KR|\theta_{kr}| + R|\phi_r|$, does not grow either with n or p. As a direct consequence, the parameter estimation of SPARTACO, conditioning on the clustering labels $\mathscr{Z}$ and $\mathscr{W}$, remains much less computationally prohibitive than the one of the MVNB, specially when the sample size becomes considerably large.

Latent Block Model is a vast class of statistical models that can be seen as an extension of the mixture model for co-clustering problems. The model for continuous data (Govaert and Nadif, 2013, Chapter 5) can be written using the Matrix Variate Normal representation as $\mathrm{X}^{kr} \sim \mathscr{MVN}(\mu_{kr}1_{n_k \times p_r}, \mathbb{I}_{n_k}, \xi_{kr}\mathbb{I}_{p_r})$ and so it is based on the assumption that the data entries in a block are independent given the clustering labels (conditional independence). The intra-block model is thus a special case of SPARTACO when $\Sigma_{kr} = \mathbb{I}$ and $\tau_{kr} = 0$, for all k and r. However, the LBM is more general on the probabilitistic assumptions over the clustering variables. In fact, it assumes $\Pr(\mathscr{Z}_i = k) = \pi_k$ and $\Pr(\mathscr{W}_j = r) = \rho_r$, where $(\pi_1, \ldots, \pi_K)$ and $(\rho_1, \ldots, \rho_R)$ are probability vectors such that $\sum_{k=1}^{K} \pi_k = \sum_{r=1}^{R} \rho_r = 1$, while SPARTACO implicitly assumes that $\Pr(\mathscr{Z}_i = k) = 1/K$ and $\Pr(\mathscr{W}_j = r) = 1/R$ for any k and r.

Supplementary Figure 1 (Sottosanti and Risso, 2022) gives a summary of the relations across SPARTACO and the co-clustering models discussed in this section.

3. Inference.

3.1. Identifiability.

The model as expressed in Formula (1) is not identifiable in the covariance term: in fact, for any $a > 0$, $\sigma_{kr,i}^2 \cdot \Delta_{kr} = a\sigma_{kr,i}^2 \cdot \Delta_{kr}/a = \tilde{\sigma}_{kr,i}^2 \cdot \widetilde{\Delta}_{kr}$. This issue generates in practice an infinite number of solutions for the parameter estimate.

A typical workaround to get unique parameter estimates consists in setting the value of some covariance parameters. In our model, this would mean taking $\sigma_{kr,i}^2 = c$, for one i in $\{1, \ldots, n_k\}$, using an arbitrary positive constant c. Incidentally, this is equivalent to constraint $\mathrm{tr}(\Sigma_{kr})$, the trace of the matrix Σ_{kr} (Allen and Tibshirani, 2010; Caponera et al., 2017). However, we discard this solution as, under our model, the rows of the data matrix are involved into a clustering procedure. Thus, it is not possible to define which i in a cluster should take the constraint.

The solution we adopt for our model puts the identification constraint on Δ_{kr} (Anderlucci and Viroli, 2015). Since $\mathrm{tr}(\Delta_{kr}) = p_r(\tau_{kr} + \xi_{kr})$, we constraint the quantity $\tau_{kr} + \xi_{kr} = c_\Delta$, where c_Δ is an arbitrary positive constant. Such constraint has a notable practical consequence: in fact, once the estimate $\hat{\tau}_{kr}$ is determined within the constrained domain $(0, c_{\Delta})$, then $\hat{\xi}_{kr}$ is simply taken by difference as $\hat{\xi}_{kr} = c_{\Delta} - \hat{\tau}_{kr}$. Hence, we can only interpret $\hat{\tau}_{kr}$ and $\hat{\xi}_{kr}$ in relation to each other and not in absolute terms. According to Svensson, Teichmann and Stegle (2018), in our applications (Sections 4 and 5) we will consider the quantity τ_{kr}/ξ_{kr} that we called *spatial signal-to-noise ratio*. This ratio is easily interpretable because it represents the amount of spatial expression of the genes in a cluster with respect to the nugget effect.

3.2. Model estimation.

To estimate SpaRTaCo, we propose an approach based on the maximization of the *classification log-likelihood*, that is

$$\log \mathscr{L}(\Theta, \mathscr{Z}, \mathscr{W}) = \sum_{i=1}^{n} \sum_{k=1}^{K} 1(\mathscr{Z}_i = k) \left\{ \sum_{r=1}^{R} \log f(x_{i.}^{\cdot r}; \theta_{kr}, \phi_r) \right\}, \tag{4}$$

where $\Theta = \cup_r \{ \cup_k \theta_{kr}, \phi_r \}$, $x_{i.}^{\cdot r}$ is the i-th row of the matrix $X^{\cdot r}$ and $f(\cdot ; \cdot)$ is given in Formula (3). Notice that the correlation across the columns does not allow to write the $\mathscr{W}$ explicitly. This issue does not concern the $\mathscr{Z}$, because the rows are independent.

Chapter 2 of Bouveyron et al. (2019) makes a clear distinction between the classification and the *complete log-likelihood* (the latter includes an additional part related to the distribution of the clustering labels). However, since SpaRTaCo implicitly assumes that $\Pr(\mathscr{Z}_i = k) = 1/K$ and $\Pr(\mathscr{W}_j = r) = 1/R$ for any k and r, then there is no practical difference between classification and complete log-likelihood.

The classification log-likelihood can be maximized with a *classification EM* algorithm (CEM, Celeux and Govaert, 1992), a modification of the standard EM which allocates the observations into the clusters during the estimation procedure. The CEM is an iterative algorithm which alternates between a classification step (CE Step), where the estimates of $\mathscr{Z}$ and $\mathscr{W}$ are updated, and a maximization step (M Step), which updates the parameter estimates of Θ. The benefits brought by such algorithm are particularly visible when complex models as the LBM are employed, because the joint conditional distribution $p(\mathscr{Z}, \mathscr{W} | X; \Theta)$ is not directly available (Govaert and Nadif, 2013).

Under SpaRTaCo, a direct update of $\mathscr{W}$ through a CE step is unfeasible due to the correlation across the columns, and so the estimation algorithm requires some modifications. This issue was already discussed by Tan and Witten (2014) for their MVNb model; however, their solution consists in an heuristic estimation algorithm with no guarantees of convergence. We propose to perform a stochastic allocation (SE step), where the column clustering configuration $\mathscr{W}$ is sampled from a Markov chain whose limit distribution is the conditional distribution $p(\mathscr{W} | \mathscr{Z}, X; \Theta)$. This step can be performed using the Metropolis-Hastings algorithm. A stochastic version of the EM algorithm was previously employed also for estimating the LBM by Keribin et al. (2015), Bouveyron et al. (2018) and Casa et al. (2021). Because of the alternation of a classification move, a stochastic allocation move and a maximization move, we name our algorithm *classification-stochastic EM* (CS-EM). We denote with $(\Theta, \mathscr{Z}, \mathscr{W})^{(t-1)}$ the estimate of the model parameters and of the clustering labels at iteration $t-1$. At step t, the algorithm executes the following steps:

- **CE Step**: keeping fixed $(\mathscr{W}, \Theta)^{(t-1)}$, update the row clustering labels with the following rule:

$$\mathscr{Z}_i^{(t)} = \underset{k=1,\dots,K}{\operatorname{argmax}} \frac{\prod_{r=1}^{R} f\left(x_{i.}^{r}; \theta_{kr}^{(t-1)}, \phi_r^{(t-1)}\right)}{\sum_{k'=1}^{K}\left\{\prod_{r=1}^{R} f\left(x_{i.}^{r}; \theta_{k'r}^{(t-1)}, \phi_r^{(t-1)}\right)\right\}}, \quad i = 1, \dots, n.$$

- **SE Step**: keeping fixed $\mathscr{Z}^{(t)}$ and $\Theta^{(t-1)}$, generate a candidate clustering configuration $\mathscr{W}^*$ by randomly changing some elements from the starting configuration $\mathscr{W}^{(t-1)}$. Let m be the number of elements of $\mathscr{W}^{(t-1)}$ that we attempt to change: m can be either fixed or randomly drawn from a discrete distribution. To formulate $\mathscr{W}^*$, we exploit two moves.

  (M1) Two clustering labels $g_1 \sim \mathscr{U}(\{1,\dots,R\})$ and $g_2 \sim \mathscr{U}(\{1,\dots,R\} \setminus \{g_1\})$ are drawn. The candidate configuration $\mathscr{W}^*$ is made by selecting m observations from $\mathscr{W}^{(t-1)}$ at random with label g_1 and changing their label to g_2. The quantity

$$\frac{q\left(\mathscr{W}^{(t-1)} | \mathscr{W}^*\right)}{q\left(\mathscr{W}^* | \mathscr{W}^{(t-1)}\right)} = \frac{p_{g_1}! p_{g_2}!}{(p_{g_1} - m)!(p_{g_2} + m)!}$$

  is the ratio of transition probabilities employed by the Metropolis-Hastings algorithm to evaluate $\mathscr{W}^*$, where $q\left(\mathscr{W}^* | \mathscr{W}^{(t-1)}\right)$ and $q\left(\mathscr{W}^{(t-1)} | \mathscr{W}^*\right)$ are respectively the probabilities of passing from configuration $\mathscr{W}^{(t-1)}$ to $\mathscr{W}^*$ and *vice-versa*. This move almost coincides with the (M2) move of Nobile and Fearnside (2007).

  (M2) For $h = 1,\dots,m$, the clustering $g_{1h} \sim \mathscr{U}(\{1,\dots,R\})$ and $g_{2h} \sim \mathscr{U}(\{1,\dots,R\} \setminus \{g_{1h}\})$ are drawn. Let $b_{lr} = \sum_{h=1}^{m} 1(g_{lh} = r)$, for $l = 1, 2$ and $r = 1, \dots, R$. Then the candidate configuration $\mathscr{W}^*$ is made by changing the labels of b_{1r} observations selected at random from the group r, when $b_{1r} > 0$, to $g_{2_{\kappa(r)}}$, where $\kappa(r) = \{h = 1,\dots,m : g_{1h} = r\}$. The ratio of transition probabilities is

$$\frac{q\left(\mathscr{W}^{(t-1)} | \mathscr{W}^*\right)}{q\left(\mathscr{W}^* | \mathscr{W}^{(t-1)}\right)} = \prod_{r: b_{2r} > 0} \frac{b_{2r}!(p_r - b_{1r})!}{(p_r - b_{1r} + b_{2r})!} \Big/ \prod_{r: b_{1r} > 0} \frac{b_{1r}!(p_r - b_{1r})!}{p_r!}.$$

  The choice between (M1) and (M2) is random. The candidate configuration $\mathscr{W}^*$ is accepted with probability $\min\{1, A\}$, where A is the following Metropolis-Hastings ratio:

$$A = \frac{\mathscr{L}\left(\Theta^{(t-1)}, \mathscr{Z}^{(t)}, \mathscr{W}^{*}\right)}{\mathscr{L}\left(\Theta^{(t-1)}, \mathscr{Z}^{(t)}, \mathscr{W}^{(t-1)}\right)} \frac{q\left(\mathscr{W}^{(t-1)} | \mathscr{W}^{*}\right)}{q\left(\mathscr{W}^{*} | \mathscr{W}^{(t-1)}\right)}.$$

Within the same iteration t, the SE Step can be run for an arbitrary large number of times to accelerate the exploration of the space of clustering configurations and so the convergence of the estimation algorithm to a stationary point. From our experience, we suggest to repeat the SE Step for at least 100 times per iteration.

- **M Step**: using the rows in $\mathscr{C}_k^{(t)}$ and the columns in $\mathscr{D}_r^{(t)}$, update the parameter estimates $\theta_{kr}^{(t)}$ and $\phi_r^{(t)}$. The derivative of the log-likelihood with respect to (θ_{kr}, ϕ_r) does not lead to closed solutions for updating the model parameters, and for this reason a numerical optimizer must be applied. We exploit the L-BFGS-B algorithm of Byrd et al. (1995) implemented in the `stats` library of the `R` computing language, which allows constrained optimization; this aspect is particularly useful to estimate τ_{kr} under the identifiability constraint described in Section 3.1.

Following Tan and Witten (2014), our implementation of the estimation algorithm alternates each allocation step, either the CE Step and the SE Step, with an M Step. As pointed by Keribin et al. (2015), the SE Step is not guaranteed to increase the classification log-likelihood at each iteration, but it generates an irreducible Markov chain with a unique stationary distribution which is expected to be concentrated around the maximum likelihood parameter estimate. The estimation algorithm must be run for a sufficiently large number of iterations. We additionally implemented a convergence criterion that stops the algorithm if the increment of the classification log-likelihood is smaller than a certain threshold for a given number of iterations in a row. The final estimates of ($\widehat{\Theta}$, $\widehat{\mathscr{Z}}$, $\widehat{\mathscr{W}}$) are the values obtained at the iteration from which (4) is maximum.

Notice that the criterion to form the co-clusters that SpaRTaCo uses has also a geometrical interpretation; in fact, in the same way that *k-means* minimizes the Euclidean distance between the observations and the centroids, SpaRTaCo minimizes the Mahalanobis distance of the observations from the block centroids, embedding the spatial structure of the data into the covariance matrix. Therefore, even when the data do not fully respect the probabilistic assumptions, the model is still valid, as a distance-based clustering algorithm.

3.3. Measuring the clustering uncertainty.

The proposed estimation procedure should be run multiple times from different starting points to check if the algorithm encounters some local maxima. In addition, the parallel runs can be used to quantify the uncertainty of the estimated co-clustering structure. In fact, if the analyzed data carry large evidence in favor of a unique clustering configuration, then the parallel runs will return approximately the same row and column clusters. If instead the clustering structure of the data that SpaRTaCo searches for is not evident, then the multiple runs of the algorithm will tend to discover different but equally likely solutions.

Let us suppose to run the CS-EM algorithm S times on the same dataset: $(\widehat{\Theta}^{(s)}, \widehat{\mathscr{Z}}^{(s)}, \widehat{\mathscr{W}}^{(s)})$ is the solution to the parameter estimate returned by the s-th run, for $s = 1, \ldots, S$, and $\ell^{(s)} = \log \mathscr{L}\left(\widehat{\Theta}^{(s)}, \widehat{\mathscr{Z}}^{(s)}, \widehat{\mathscr{W}}^{(s)}\right)$. In addition, Let $s^* = \mathrm{argmax}_s\, \ell^{(s)}$: since the co-clustering structure $(\widehat{\mathscr{Z}}^{(s^*)}, \widehat{\mathscr{W}}^{(s^*)})$ has found the largest evidence across the S runs on the current data, it is the final estimate returned by the algorithm. The co-clustering uncertainty can be thought of as a function of the distances between the final estimate, $(\widehat{\mathscr{Z}}^{(s^*)}, \widehat{\mathscr{W}}^{(s^*)})$, and the other estimates of lower evidence, $(\widehat{\mathscr{Z}}^{(s)}, \widehat{\mathscr{W}}^{(s)})$, for $s \neq s^*$. Let $\mathscr{I}_k = \left\{1\left(\widehat{\mathscr{Z}}_i^{(s^*)} = k\right)\right\}_{1 \le i \le n}$ be the binary vector denoting which rows belong to the k-th row cluster given by the run s^*, for $k = 1, \ldots, K$, and $\mathscr{I}_{h_s(k)} = [1\{\widehat{\mathscr{Z}}_i^{(s)} = h_s(k)]_{1 \le i \le n}$ be the binary vector denoting which observations belong to the cluster $h_s(k)$ given by the s-th run, where $h_s(k) = \mathrm{argmax}_{h = 1,\ldots,K} \sum_{i=1}^{n} 1\left(\mathscr{Z}_i^{(s^*)} = k, \mathscr{Z}_i^{(s)} = h\right)$ and $s \neq s^*$. In addition, let us consider the weights $\omega_s = 1/\left(\ell^{(s^*)} - \ell^{(s)}\right)$. The uncertainty of the row cluster k is measured as

$$\varepsilon_k^{\text{rows}} = \frac{\sum_{s \neq s^*} \omega_s \mathrm{CER}(\mathscr{I}_k, \mathscr{I}_{h_s(k)})}{\sum_{s \neq s^*} \omega_s}, \tag{5}$$

where $\mathrm{CER}(\cdot, \cdot)$ denotes the *clustering error rate* (Witten and Tibshirani, 2010), an index that measures the disagreement between a reference and an estimated clustering configuration: the closer is CER to 0, the larger is the agreement between the true and the estimated clusters. The $\{\omega_s\}_{s \neq s^*}$ give a large weight to the CER between $\mathscr{I}_k$ and $\mathscr{I}_{h_s(k)}$ when $\ell^{(s^*)} - \ell^{(s)}$ is small, and *vice-versa*. The reason is intuitively that, if both ω_s and $\mathrm{CER}(\mathscr{I}_k, \mathscr{I}_{h_s(k)})$ are large, then there are two considerably different clustering configurations that yield approximately the same log-likelihood value. Thus, the clustering structure of the data is uncertain. If instead ω_s is small, the difference between $\widehat{\mathscr{Z}}^{(s^*)}$ and $\widehat{\mathscr{Z}}^{(s)}$ is in practice irrelevant, because the evidence arising from the data clearly leans in favor of $\widehat{\mathscr{Z}}^{(s^*)}$.

Formula (5) can be applied also for computing the uncertainties of the column clusters $(\varepsilon_1^{\text{cols}}, \ldots, \varepsilon_R^{\text{cols}})$, just replacing $\widehat{\mathscr{Z}}^{(s)}$ with $\widehat{\mathscr{W}}^{(s)}$. The uncertainty measure introduced here can be interpreted similarly to the CER index: the closer are $\varepsilon_k^{\text{rows}}$ and $\varepsilon_r^{\text{cols}}$ to 0, the larger is the evidence of a unique co-clustering structure of the data.

3.4. Model selection.

SpaRTaCo can be run with different spatial covariance models $k(\cdot\,;\cdot)$ and with different combinations of K and R. We consider the problem of selecting the best model for the data, both in terms of the number of clusters and the spatial covariance function, using an information criterion. The most common criteria, the AIC and the BIC, cannot be derived under Model (1) because the likelihood of the data $p(\mathrm{X}; \Theta)$, marginalized with respect to the latent variables $\mathscr{Z}$ and $\mathscr{W}$, is not available in closed form.

In this work, we propose to guide the model selection using the *integrated completed log-likelihood* (ICL, Biernacki, Celeux and Govaert, 2000). The ICL is a well-established criterion for selecting the number of clusters (Bouveyron et al., 2019) which has become popular in the co-clustering framework for selecting the size of LBM (Keribin et al., 2015; Bouveyron et al., 2018; Casa et al., 2021). Under Model (1)-(2), its expression is

$$\mathrm{ICL} = \log \mathscr{L}\left(\widehat{\Theta}, \widehat{\mathscr{Z}}, \widehat{\mathscr{W}}\right) - n \log K - p \log R - \frac{4KR + \dim(\phi)R}{2} \log np, \tag{6}$$

where $\dim(\phi)$ is the dimension of the parameter vector ϕ_r , which does not depend on r. The derivation of (6) is described more in details in Supplementary Section 1. Operatively, the best model from a list of candidates corresponds to the one with the largest value of (6).

In the presence of mixed effects, Delattre et al. (2014) argue that the actual sample size is not trivial to define, and thus the classical information criteria need to be modified. In particular, they derive an alternative formulation of the BIC which includes a term that depends only on the parameters involved with the random effects. However, their model specification assumes that the marginal distribution of the data with the random parameters integrated out cannot be derived in closed form. Although the presence of the random variances $\sigma^2_{kr,i}$ makes SpaRTaCo a random effect model, the integration of $\sigma^2_{kr,i}$ from the density function of $x^{kr}_{i.} | \sigma^2_{kr,i}$ leads to the marginal density (3). For this reason, we do not implement any modification based on the random effects into our information criterion (6).

4. Simulation studies.

4.1. Simulation model.

We study the performance of SpaRTaCo with five simulated spatial experiments that recreate some possible scenarios that can be found in real data. We generate the latent blocks using the matrix-variate normal distribution (Gupta and Nagar, 2018) as follows: given the number of row and column clusters K^{true} and R^{true} (for convenience, we considered here $K^{\text{true}} = R^{\text{true}} = 3$ in every simulation experiment), the clustering labels $\mathscr{Z}^{\text{true}}$ and $\mathscr{W}^{\text{true}}$, and the clusters $\mathscr{C}_k^{\text{true}} = \{i = 1, \ldots, n : \mathscr{Z}_i^{\text{true}} = k\}$ and $\mathscr{D}_r^{\text{true}} = \{j = 1, \ldots, p : \mathscr{W}_j^{\text{true}} = r\}$, the (k, r)-th block is drawn from

$$\mathrm{X}^{kr} \sim \mathscr{MVN}\left(\mu_{kr}^{\text{true}} 1_{n_k \times p_r}, \Sigma_{kr}^{\text{true}}, \Delta_{kr}^{\text{true}}\right), \quad \Delta_{kr}^{\text{true}} = \tau_{kr}^{\text{true}} \mathscr{K}_r^{\text{true}}\left(S^r; \phi_r^{\text{true}}\right) + \xi_{kr}^{\text{true}} \mathbb{I}_{p_r}, \tag{7}$$

where $\mathscr{K}_r^{\text{true}}\left(S^r; \phi_r\right) = \left(k_r^{\text{true}}\left(\|s_j^r - s_{j'}^r\|; \phi_r^{\text{true}}\right)\right)_{1 \le j, j' \le p_r}$, and $k_r^{\text{true}}(\,\cdot\,; \phi_r^{\text{true}})$ is an isotropic spatial covariance kernel parametrized by ϕ_r^{true}. Note that, differently from (2), the presence of the subscript r into the kernel matrix $\mathscr{K}_r^{\text{true}}$ denotes that the spatial covariance function can be different for any column cluster. In our simulations, we employed the *Exponential* kernel with scale θ_E for the columns in $\mathscr{D}_1^{\text{true}}$, the *Rational Quadratic* kernel with parameters (θ_R, α_R) for the columns in $\mathscr{D}_2^{\text{true}}$, and the *Gaussian* kernel (known also as *Squared Exponential*) with scale θ_G for the columns in $\mathscr{D}_3^{\text{true}}$. Their formulation is reported in Supplementary Section 2 and it is further discussed in Chapter 4 or Rasmussen and Williams (2006). The simulation model (7) implies the following marginal distributions of the genes and of the spots:

$$\mathrm{x}_{i.}^{k.} | \mathscr{Z}^{\text{true}}, \mathscr{W}^{\text{true}} \sim \mathscr{N}_p\left\{\left(\mu_{k1}^{\text{true}} 1_{p1}, \ldots, \mu_{k3}^{\text{true}} 1_{p3}\right), \Sigma_{ii}^{\text{true}} \operatorname{diag}\left(\Delta_{kr}^{\text{true}}\right)_{r=1,2,3}\right\}, \tag{8}$$

$$\mathrm{x}_{.j}^{.r} | \mathscr{Z}^{\text{true}}, \mathscr{W}^{\text{true}} \sim \mathscr{N}_n\left\{\left(\mu_{1r}^{\text{true}} 1_{n1}, \ldots, \mu_{3r}^{\text{true}} 1_{n3}\right), \ c^{\text{true}} \operatorname{diag}\left(\Sigma_{k}^{\text{true}}\right)_{k=1,2,3}\right\}, \tag{9}$$

where $\Sigma_{ii}^{\text{true}}$ is the variance parameter of the i-th row and does not depend on k, and the notation $\operatorname{diag}(\Delta_{kr}^{\text{true}})_{r=1,2,3}$ denotes a block diagonal matrix formed by the matrices $\Delta_1^{\text{true}}, \ldots, \Delta_3^{\text{true}}$. Notice that, from Formula (9), the marginal distribution of the spots does not carry any information on the column clusters. The cross-covariance matrix of two rows $i, i' \in \mathscr{C}_k^{\text{true}}$ is $\operatorname{Cov}\left(\mathrm{x}_{i.}^{k.}, \mathrm{x}_{i'}^{k}\right) = \Sigma_{k,ii'}^{\text{true}} \operatorname{diag}\left(\Delta_{kr}^{\text{true}}\right)_{r=1,2,3}$, and the cross-covariance of two columns $j, j' \in \mathscr{D}_r^{\text{true}}$ is $\operatorname{Cov}(x_{.j}^{.r}, \ x_{.j'}^{.r}) = \operatorname{diag}\{\tau_{kr}^{\text{true}} k_r^{\text{true}}(s_j^r - s_{j'}^r; \ \phi_r^{\text{true}}) \Sigma_k^{\text{true}}\}_{k=1,2,3}$.

We took the sets of spatial coordinates $(\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3)$ from the brain tissue sample of the subject with ID 151507 contained in the R package `spatialLIBD` and processed with Visium. As we briefly discussed in Section 1.1, the spots in these experiments have been manually annotated into layers. We extracted 200 spots from each of the three layers appearing in the top-right region of the image. The resulting map of 600 spots is shown in the left plot of Figure 3; the clustering labels $\mathscr{W}^{\text{true}}$ correspond to the labels assigned with the manual annotation. Note that, although we took the spot annotation from the real data, the image clusters in the simulation experiments have a substantially different meaning: in fact, under the simulation model (7), they denote regions of the tissue in which some genes are expressed with specific spatial variation profiles, while, in the real data, the manually annotated regions identify the morphological structure of the tissue. In addition, the right plot of Figure 3 shows the covariance functions used for the simulations. We set the covariance parameters $(\theta_E, \theta_R, \alpha_R, \theta_G)$ according to how much the clusters extend over the plane: the covariance function of $\mathscr{D}_1^{\text{true}}$ is steeper than the one of $\mathscr{D}_2^{\text{true}}$ because $\mathscr{D}_1^{\text{true}}$ covers a smaller distance. Because $\mathscr{D}_3^{\text{true}}$ is made of two distinct groups of spots appearing in the top and in the bottom of Figure 3 (left), we specify $k_3^{\text{true}}(\cdot \ ; \ \cdot)$ in such a way that only the spots within the same group are spatially correlated, while spots from different groups are poorly correlated. Details on the covariance parameters are given in the caption of Figure 3.

Last, we set the values of the spatial signal-to-noise ratios $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}}$. The additional identifiability constraint $\tau_{kr}^{\text{true}} + \xi_{kr}^{\text{true}} = c_{kr}^{\text{true}}$ leads to a unique value of the parameters τ_{kr}^{true} and ξ_{kr}^{true}. Note that, due to the identifiability issue described in Section 3.1, which holds also for the simulation model, the value assigned to c_{kr}^{true} is in practice irrelevant. For this reason, without loss of generality we assumed $c_{kr}^{\text{true}} = c^{\text{true}} = 10$, for any k and r. In our simulations, we considered three cases: (i) no spatial effect, $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}} = 0$; (ii) the spatial effect is as much as the nugget effect, $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}} = 1$; and (iii) the spatial effect is considerably larger than the nugget effect, $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}} = 3$. Finally, we set $\mu_{kr}^{\text{true}} = 0$ to test if SpaRTaCo is able to recover the co-clusters using the covariance of the data without being driven by the effect of the mean.

4.2. Competing models and evaluation criteria.

We fit SpaRTaCo on the simulated data taking $k(\cdot;\ \cdot)$ in Formula (2) as the exponential kernel, which has a lower decay than the more common Gaussian kernel considered by Svensson, Teichmann and Stegle (2018) and Sun, Zhu and Zhou (2020). The estimation is carried running the algorithm described in Section 3.2 five times in parallel to avoid local maxima. The procedure is run for 5,000 iterations, and if the classification log-likelihood function is still growing, it is run until reaching 10,000 iterations. In addition to SpaRTaCo, we consider also the following co-clustering models:

- two independent k-means, applied separately to the rows and to the columns of the data matrix, using the `R` function `kmeans`;
- the biclustering algorithm BC, and its sparse version sparseBC with $\lambda = 1$, 10, 20, using the `R` package sparseBC;
- the matrix variate normal algorithm MVNb with the following setups: *1)* $\lambda = 1$, $\rho\Sigma = \rho\Delta = 0.25$, *2)* $\lambda = 10$, $\rho\Sigma = \rho\Delta = 2.5$ and *3)* $\lambda = 20$, $\rho\Sigma = \rho\Delta = 5$. We had to implement a slight modification of the function `matrixBC` contained in the `R` package `sparseBC`, as its original form could not handle the computation of the logarithm of the determinant of some matrices.
- LBM, using the `R` package `blockcluster`;

Tan and Witten (2014) do not give any indication on how to select the penalization parameters $\rho\Sigma$ and $\rho\Delta$ of MVNb. In their simulation experiments and real data applications, they simply set λ to be much larger than $\rho\Sigma$ and $\rho\Delta$. For this reason, in our simulations we fit MVNb with three setups, where the λ values are the same of sparseBC, and $\rho\Sigma$ and $\rho\Delta$ are taken equal to a quarter of λ. We measure the clustering accuracy by comparing the estimated row and column clusters with the true ones using the CER. In this section, we do not focus on the parameter estimates returned by SpaRTaCo, because the principal goal is evaluating the classification accuracy of the models. We leave the interpretation of the parameter estimates to Section 5.

4.3. Simulation 1.

We generated 9 blocks of size $n_k = 200 \times p_r = 200$, for every k and r. We assume that the variances and covariances of the genes do not change with respect to the spot clusters, thus $\Sigma_{kr}^{\text{true}} = \Sigma_k^{\text{true}}$ for all r. We draw Σ_k^{true} as follows:

$$\Sigma_1^{\text{true}} \sim \mathscr{W}i(210,\ 0.03\mathbb{I}_{200}), \quad \Sigma_2^{\text{true}} \sim \mathscr{W}i(230,\ 0.05\mathbb{I}_{200}), \quad \Sigma_3^{\text{true}} \sim \mathscr{W}i(200,\ \Sigma_1^{\text{true}}/150), \tag{10}$$

where $\mathscr{W}i(a,\ b)$ denotes a Wishart distribution with degrees of freedom a and scale matrix b. Generating the covariance matrices from a Wishart distribution ensures that the draws are positive definite. The simulation setup in Formula (10) was selected after both numerical and graphical evaluations. More details on the motivations which led to this setup are given in Supplementary Section 3.

We designed a spatial experiment in which three clusters of genes have a grade of spatial expression that changes in three different areas of the tissue sample. The tessellation of the data matrix into blocks and the values of the spatial signal-to-noise ratios appear in Figure 4 (a). Figure 5 displays a spatial experiment generated under this framework, to show how the average gene expression changes across the 9 blocks. For example, in the left panel $(k = 1)$ there is an evident spatial expression across the spots from clusters $r = 2$ and $r = 3$, while the spots in $r = 1$ are randomly positive or negative due to the absence of spatial dependency. Different spatial expression profiles across the image are distinguishable also in real data, as seen in Supplementary Figure 3, which displays the expression of three genes on the subject 151507. The real and simulated experiments appear very similar, confirming that our simulations are realistic and can be used for testing methods designed for 10X Visium data. We simulated 10 replicates of this experiment and we fitted the co-clustering models listed in Section 4.2 using $K = R = 3$. The boxplots of the row and the column CER over the 10 replicates appear in the first line of Figure 6. SpaRTaCo outperforms the competing models and leads to no clustering errors. Good results on the rows are achieved also by the LBM, while on the columns the k-means type algorithms (k-means, BC and sparseBC) and the MVNb with $\rho\Sigma = \rho\Delta = 5$ perform better than the other competitors. A further confirmation of the accuracy of SpaRTaCo for modelling this spatial experiment comes from the value of estimated clustering uncertainties, which are $\varepsilon_k^{\text{rows}} < 0.001$ and $\varepsilon_r^{\text{cols}} < 0.001$, for every k and r. A graphical representation of these quantities across the 10 replicates is given in Supplementary Figure 4.

This experiment has demonstrated that the presence of spatial covariance patterns, if not properly accounted for, heavily impacts on the performance of the standard co-clustering models. Since the MVNb is designed to flexibly estimate the covariance of the blocks, in theory it should be the best candidate for such complex experiments. However, the formulation of $\widehat{\Sigma}_k^{\text{MVNB}}$ and $\widehat{\Delta}_r^{\text{MVNB}}$ is too generic for capturing the spatial correlation across the spots, causing a poor clustering result. As a confirmation of this statement, we notice that the smallest classification error made by MVNb is reached when the penalization parameters $\rho\Sigma$ and $\rho\Delta$ are large, leading the estimated matrices $\widehat{\Sigma}_k^{\text{MVNB}}$ and $\widehat{\Delta}_r^{\text{MVNB}}$ to be diagonal.

As a second step of this experiment, we tested the model selection criterion based on the ICL proposed in Section 3.4. Using the same 10 replicates of the experiment, we ran SpaRTaCo with K and R taking values in $\{2, 3, 4\}$. Supplementary Figure 5 shows that the proposed ICL always selects the correct model dimension, while the classification log-likelihood favors models with a larger number of co-clusters than the truth.

While the ICL criterion accurately selects the number of co-clusters, it is a computationally expensive procedure due to the large number of times that the estimation must be run. Hence, we compared our model selection method with two faster alternatives: the first selects the number of row and column clusters separately by combing a dimension reduction method with k-means (details are given in Supplementary Section 4), the second, proposed by Tan and Witten (2014), performs a 10-fold cross-validation using sparseBC; a function that implements this last method can be found into the R package sparseBC. The first criterion selected 6 row clusters on the 90% of the replicates of Simulation 1, and 5 clusters in the remaining 10%; on the columns, it selected 3 clusters on the 33% of the replicates, and 4 clusters on the remaining 77%. The second criterion was applied with K and R taking values in $\{2, \ldots, 6\}$ and fixing $\lambda = 10$, but it has revealed to be inadequate for this kind of data, as it selected $K = 6$ and $R = 6$ on every replicate of the experiment.

4.4. Simulation 2.

The second simulation experiment differs from the first in the values of the spatial signal-to-noise ratios, which are now taken as in Figure 4 (b). For any r, the signal-to-noise ratios $\left\{\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}},\ k = 1, \ldots, K^{\text{true}}\right\}$ have all the same value. As a consequence, $\Delta_{kr}^{\text{true}} = \Delta_{r}^{\text{true}}$ for any k. Under the current setup, the marginal distribution of a row $i \in \mathscr{C}_k^{\text{true}}$ under the data generating model given in Formula (8) does not depend on k and so it is not informative of the row clustering. The only discriminating facets are the cross-covariances of the rows and of the columns, which carry the information about the row clusters through the matrices Σ_k^{true}. This framework is thus meant to evaluate the performance of SpaRTaCo when all the genes have the same spatial expression profiles across the tissue. A representation of a spatial experiment generated under this framework is given in the top row of Supplementary Figure 6.

We ran the co-clustering models using $K = R = 3$ on 10 replicates on the proposed experiment; the results are displayed in the second line of Figure 6. Our model outperforms the competitors: on the rows, the median CER from SpaRTaCo is less than 0.2, while on the columns it returns a perfect classification on all replicates. The estimated row clustering uncertainty is low ($\varepsilon_k^{\text{rows}} < 0.15, \forall k$), while the column clustering uncertainty is practically null. Details are given in the second row of Supplementary Figure 4. Both Simulations 1 and 2 have shown that SpaRTaCo works properly even if the spatial covariance function employed by the fitted model in Formula (2) does not match the covariance functions of the data generating process. In particular, Simulation 2 has highlighted this remarkable result because the only cluster of columns for which the spatial covariance function is correctly specified is $r = 1$, which however is devoid of any spatial effect, as $\tau_{k1}^{\text{true}} = 0$ for any k.

The best competitor on the rows is the LBM, with a median CER of 0.44. On the columns, the best results are from the k-means type models, or alternatively from the MVNb with $\lambda = 20$ and $\rho\Sigma = \rho\Delta = 5$. Considerable results are obtained also with the LBM; however, its classification accuracy is more variable. This experiment hence confirms what we have already observed in Simulation 1, namely that, in the presence of spatial covariance patterns in the data, the model of Tan and Witten (2014) tends to fail in recovering the correlation structure, at least in our simulation setup. This is demonstrated by the diagonal estimated covariance matrices $\left\{\widehat{\Sigma}_k^{\text{MVNB}},\ k = 1,\ 2,\ 3\right\}$ and $\left\{\widehat{\Delta}_r^{\text{MVNB}},\ r = 1,\ 2, 3\right\}$.

4.5. Simulation 3.

The third simulation experiment assumes that the spatial signal-to-noise ratio $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}}$ is constant across the blocks within the same row cluster k; as a consequence, $\tau_{kr} = \tau_k$ for any r. This case is illustrated in Figure 4 (c). Notice for example that the rows in $\mathscr{C}_1^{\text{true}}$ are not spatially expressed in any of the three column clusters. Under the current simulation setup, the marginal distribution of a row $i \in \mathscr{C}_k^{\text{true}}$ given in Formula (8) is informative on the column clusters only through the different spatial kernels $k_r^{\text{true}}(\,\cdot\,;\,\phi_r^{\text{true}})$, while, as already discussed in Section 4.1, the marginal distribution of a column $j \in \mathscr{D}_r^{\text{true}}$ is never informative on the column clusters. The cross-covariances of the rows and of the columns are informative of both rows and column clustering. Under this framework, it is challenging to determine the image areas with spatial interaction, because all the genes in a cluster $\mathscr{C}_k^{\text{true}}$ are spatially expressed with the same intensity over the whole tissue. An example of a spatial experiment generated under this simulation setup is given in the bottom row of Supplementary Figure 6.

We ran the co-clustering models on 10 replicates of the experiment using $K = R = 3$; the results appear in the third line of Figure 6. On the rows, SpaRTaCo outperforms the competitor models returning a CER of zero for all replicates. On the columns, its clustering accuracy is highly variable: the median CER is 0.21, the first and the third quartiles are 0.08 and 0.25, and extremes are 0 and 0.36. The competitor models, and in particular the k-means type models, are substantially less variable than SpaRTaCo. Their median column CER is 0.13. However, none of them ever returns a perfect classification.

Even if SpaRTaCo has returned unstable results on the columns, the advantages brought by our model against the competitors are many, and are particularly visible from the results on the rows. The column clustering changes considerably across the replicates because, in the current setup, our estimation algorithm is more sensible to the starting points. This aspect is highlighted also by the estimated column clustering uncertainties $\varepsilon_r^{\text{cols}}$, whose values across the 10 replicates are now mainly between 0.3 and 0.4 (see Supplementary Figure 4). From our experience, if independent runs of the estimation algorithm reach distant stationary points, both the number of starting points and the number of iterations of the SE Step should be increased to favor a faster exploration of the space of the configurations.

4.6. Simulation 4.

Up to now, we built the simulation experiments under the framework in which SpaRTaCo is designed to work properly, that is the case where the genes/spots in a cluster are correlated only with the other genes/spots of the same cluster. In this section, we violate this assumption and we design a spatial experiment where both the genes and the spots are correlated also with genes and spots from other clusters. This experiment aims to study the effects of an additional dependency structure across the data that is not accounted by the fitted model.

Let X_s be a 600×600 spatial experiment made of 9 equally sized blocks, generated as in Simulation 1, and $\mathrm{X}_b \sim \mathscr{MVN}(0,\, \Sigma_b,\, \Delta_b)$. Both Σ_b and Δ_b are squared matrices of size 600: the first is drawn from $\Sigma_b \sim \mathscr{W}(600,\, 0.015\mathbb{I}_{600})$, the second is $\Delta_b = \tau_b \mathscr{K}^b(S;\, \sigma_b) + \xi_b \mathbb{I}_{600}$, where $\mathscr{K}^b(S;\, \sigma_b) = \left(k^b\left(\|s_j - s_{j'}\|;\, \sigma_b\right)\right)_{1 \le j,\, j' \le 600}$ and $k^b(\,\cdot\,;\,\sigma_b)$ is a Gaussian kernel with scale σ_b We set $\tau_b = \xi_b = c^{\text{true}}/2$ and $\sigma_b = 50$. The final simulation experiment is made as follows: $\mathrm{X} = \lambda_s \mathrm{X}_s + \lambda_b \mathrm{X}_b$, where $\lambda_s,\, \lambda_b \ge 0$. We generated 10 replicates of the current experiment, each time drawing first the matrices X_s and X_b, and then combining them to form X. Supplementary Figure 7 shows a single realization of X_s, X_b and X using $\lambda_s = \lambda_b = \sqrt{0.5}$. This value satisfies the constraint $\lambda_s^2 + \lambda_b^2 = 1$ that we imposed to keep the variance of the current experiment comparable with the previous experiments proposed in this work.

We ran the co-clustering models using $K = R = 3$; results appear in the last row of Figure 6. Despite the additional correlation structure in the data brought by the nuisance signal X_b, SpaRTaCo outperforms its competitors on both the row and the column clustering. In the right plot, the CER boxplots are more variable than in the left plot, therefore, the nuisance component has affected more the column than the row clustering of the employed models. Among the competitors, k-means and MVNb with $\lambda = 10$ and $\rho\Sigma = \rho\Delta = 2.5$ are the least affected by the nuisance: the former because it performs the clustering on the two dimensions of the data matrix separately, the latter because it regulates the estimate of the row and column covariances with a moderate shrinkage factor. The effect of the additional dependency structure is visible also on the distributions of $\varepsilon_k^{\mathrm{rows}}$ and $\varepsilon_r^{\mathrm{cols}}$, which are displayed in the last line of Supplementary Figure 4: over the 10 replicates, the row clustering uncertainties spread between 0 and 0.17, and column uncertainties between 0 and 0.5.

4.7. Simulation 5.

In the last experiment, we intentionally violate two important assumptions made by SpaRTaCo: the first states that the latent block structure of an experiment corresponds to a segmentation of the data matrix into K row clusters and R column clusters, the second states that the spatial covariance functions change only across the spots and not across the genes. For instance, we generate a spatial experiment creating first the R^{true} column clusters, and then generating the K^{true} row clusters independently for each column cluster. From a biological perspective, this setup simulates the case where the expression profiles of some genes are similar only in some specific areas of the tissue sample. In addition, following the discoveries of Svensson, Teichmann and Stegle (2018) and Sun, Zhu and Zhou (2020) that different genes are s.e. according to different spatial covariance functions, we consider a data generating model where the spatial kernels change with respect to the gene cluster index k and no longer with respect to the spot cluster index r.

Let $\mathscr{C}_{kr}^{\mathrm{true}}$ and $\mathscr{D}_r^{\mathrm{true}}$ be the actual row and column clusters, with $k = 1,\ldots, K^{\mathrm{true}}$ and $r = 1,\ldots, R^{\mathrm{true}}$, where $\mathscr{C}_{kr}^{\mathrm{true}} = \{i = 1,\ldots, n : \mathscr{Z}_{\mathrm{ir}}^{\mathrm{true}} = k\}$ is the k-th row cluster within the r-th column cluster, and $|\mathscr{C}_{kr}^{\mathrm{true}}| = n_{kr}$. Under the current setup, we draw $\mathrm{X}^{kr} \sim \mathscr{MVN}(\mu_{kr}1_{n_k \times p_r}, \Sigma_{kr}^{\mathrm{true}}, \Delta_{kr}^{\mathrm{true}})$, where the covariance across the spots is now equal to $\Delta_{kr}^{\mathrm{true}} = \tau_{kr}^{\mathrm{true}} \mathscr{K}_k^{\mathrm{true}}(S^r; \phi_k^{\mathrm{true}}) + \xi_{kr}^{\mathrm{true}} \mathbb{I}_{p_r}$. Notice that, differently from Section 4.3, the covariance matrices of the rows $\Delta_{kr}^{\mathrm{true}}$ change with respect to r because the number of observations in the cluster is n_{kr} (and no longer n_k). In addition, the model assumes that the kr-th block has mean μkr. The tessellation of the data matrix into blocks is shown in Figure 4 (d). The size of the clusters is $n_{kr} = 200$ for $k = 1, 2, 3$ and $r = 1, 2$, while $n_{13} = 100$, $n_{23} = 200$ and $n_{33} = 300$. The covariance matrices of the rows are drawn as follows:

$$\Sigma_{1r}^{\mathrm{true}} \sim \mathscr{W}i(n_{1r} + 10,\ 0.03\mathbb{I}_{n_{1r}}), \quad \Sigma_{2r}^{\mathrm{true}} \sim \mathscr{W}i(n_{2r} + 30,\ 0.05\mathbb{I}_{n_{2r}}), \quad \Sigma_{3r}^{\mathrm{true}} \sim \mathscr{W}i(n_{3r},\ \Sigma_{3r}^{*}/150),$$

where $\Sigma_{3r}^{*} \sim \mathscr{W}i\left(n_{3r} + 10,\ 0.03\mathbb{I}_{n_{3r}}\right)$. Notice that this setting is nothing but a generalization of what appears in Formula (10). Calling $\mu_{k.}^{\text{true}} = \left(\mu_{k1}^{\text{true}},\ \mu_{k2}^{\text{true}},\ \mu_{k3}^{\text{true}}\right)$, we set the mean values equal to $\mu_{1.}^{\text{true}} = (-3,\ 0,\ 3)$, $\mu_{2.}^{\text{true}} = (3,\ -3,\ 0)$ and $\mu_{3.}^{\text{true}} = (0,\ 3,\ -3)$. Finally, the employed spatial signal-to-noise ratio values $\{\tau_{kr}/\xi_{kr}\}$ are shown in Figure 4 (e).

To facilitate the model evaluation and the interpretation of the results, we assign to every row i an alternative clustering label $\mathscr{Z}_i^{*\text{true}}$ such that $\mathscr{Z}_i^{*\text{true}} = \mathscr{Z}_{i'}^{*\text{true}}$ if $i,\ i' \in \left(\mathscr{C}_{k_1 1}^{\text{true}} \cap \mathscr{C}_{k_2 2}^{\text{true}} \cap \mathscr{C}_{k_3 3}^{\text{true}}\right)$, for some $k_1,\ k_2,\ k_3 \in \{1,\ 2,\ 3\}$. In words, this means that the new clusters are formed by the rows that belong to the same cluster in all of the three column clusters. The new row clustering labels appear on the right side of Figure 4 (d). In our experiment, every $\mathscr{Z}_i^{*\text{true}} \in \{1, \ldots, 6\}$, and $\mathscr{C}_b^{*\text{true}} = \left\{i = 1, \ldots, n : \mathscr{Z}_i^{*\text{true}} = b\right\}$ is the b-th alternative cluster with size $\left|\mathscr{C}_b^{*\text{true}}\right| = 100$, for $b = 1, \ldots, 6$.

To reduce the computational cost spent on the simulation, we generated a single replicate of the experiment, and we fitted SpaRTaCo using $K = 3, \ldots,\ 9$, while the number of column clusters is kept equal to its real value, $R = 3$. Supplementary Figure 8 (a) shows that the ICL criterion selects $K = 8$ as the optimal model dimension; using the log-likelihood, we would have wrongly picked $K = 9$, confirming the importance of using a suitable information criterion to drive the model selection. In addition, one could consider also a model with a smaller number of row clusters: for example, $K = 5$ looks also a reasonable choice, because it corresponds to a local maximum. SpaRTaCo with $K = 8$ returns a row CER of 0.028 and a column CER of 0. In details, the model correctly recovers the gene clusters 2, 4, 5 and 6, while the genes in $\mathscr{C}_1^{*\text{true}}$ and $\mathscr{C}_3^{*\text{true}}$ are split into two separate groups. The estimated clustering uncertainty is $\varepsilon_r^{\text{cols}} < 0.004$, for $r = 1,\ 2,\ 3$, while on the rows it varies between 0 and 0.19. Thus, some of the genes clusters are clearly visible, while others are unstable. As a comparison, we give also the results using $K = 5$. The CER on the rows is 0.056, and it is 0 on the columns; the estimated clustering uncertainties are $\varepsilon_k^{\text{rows}} < 0.001$, for $k = 1, \ldots,\ 5$, and $\varepsilon_r^{\text{cols}} < 0.06$, for $r = 1,\ 2,\ 3$. The fact that the row clusters of the model with $K = 5$ are more stable than the ones with $K = 8$ gives additional support to the idea of selecting the model with a smaller number of blocks, but both models yield reasonably good results.

We finally run the competing models using $K = 5$ and $R = 3$; results are shown in Supplementary Figure 8 (b). Thanks to the difference in mean across the blocks, all the competing models can clearly distinguish the clustering structure of the spots. However, due to the spatial dependency effects, their performance in clustering the genes is poor, confirming, once again, the improvement brought by SpaRTaCo.

5. Application.

In this section, we analyze the human dorsolateral prefrontal cortex sample from the subject 151673 studied by Maynard et al. (2021) that we briefly described in Section 1.1 and shown in Figure 1. The dataset has 33,538 genes measured over 3,639 spots. Similarly to 10X scRNA-seq protocols, 10X Visium yields *unique molecular identifier* (UMI) counts as gene expression values. As a first step, we sought to exclude uninformative genes and reduce the analysis to a lower dimensional problem. We applied the gene selection procedure for

UMI count data proposed by Townes et al. (2019), i.e., we fit a multinomial model on every vector of gene expression and compute the deviance. Based on the criterion that large deviance values are associated to informative genes, we kept the first 500 genes and discarded the remaining ones. Supplementary Figure 9 shows that the deviance, which is very high for the top genes, reaches a plateau after 200 genes. To normalize the data, we computed, for each selected gene, the deviance residuals based on the binomial approximation of the multinomial distribution as done in Townes et al. (2019). The result of this procedure is the expression matrix X whose entries are $x_{ij} \in \mathbb{R}$ and whose row vectors $\mathrm{x}_{i\cdot}$ yield approximately symmetric histograms. Boxplots of the transformed gene expression vectors are given in Supplementary Figure 10, where it is shown also that there is no practical difference between using the binomial or the Poisson for computing the the residuals.

We fitted SpaRTaCo with all the configurations in $\{(K,\ R): K = 2,\ R = 7, \ldots, 12\}$, starting the estimation of each model from five different initial points. More details about the setup of the estimation algorithm and the computational costs are given in Supplementary Section 5. The range of column cluster values reflects the number of biological layers that appear in Figure 1. As we already mentioned in Section 1.1, SpaRTaCo performs a substantially different image clustering than BayesSpace or GIOTTO; thus, we do not expect the clusters discovered by SpaRTaCo to match the cortical layers. However, we believe that their number could still be indicative of the biological diversity of this specific area. Supplementary Figure 11 (a) gives the ICL values of the models with $K = 2$. Although our criterion selects $K = 2$ and $R = 12$, we believe that the local maximum in correspondence of $(K = 2,\ R = 9)$ represents also a valid solution. In fact, a large value of R would result in too many small clusters, complicating the biological interpretation. Furthermore, we fixed $R = 9$ and we explored the options $K \in \{1,\ 3,\ 4\}$ to investigate the absence of gene clusters $(k =\ 1)$ and the presence of multiple clusters. However, the ICL selects $K = 2$. Figure 7 (a) displays the tissue map colored according to the estimated clusters. The White Matter spots are covered by clusters $\mathscr{D}_1$, $\mathscr{D}_7$, $\mathscr{D}_8$ and $\mathscr{D}_9$; this last one is placed at the border between the White Matter and Layer 6. The remaining clusters cover the surface within the Layers 2–6. Last, Layer 1 is covered by $\mathscr{D}_4$ and mostly by $\mathscr{D}_5$ Incidentally, we note that the spot clusters within the White Matter are the ones with the smallest grade of uncertainty (see the right plot in Supplementary Figure 11 (b)).

As for the row clustering, 109 of the genes in Cluster $\mathscr{C}_2(n_2 = 129)$ were ranked within the top 200 most informative genes by the deviance procedure of Townes et al. (2019). Figure 7 (b) displays the spots colored according to $n_2^{-1}\left(\mathrm{X}^{2\cdot}\right)^T 1_{n_2}$, the average expression of the genes in $\mathscr{C}_2$, from which it emerges that the expression tends to be larger within the White Matter than in the rest of the cortical area. Panels (c) and (d) in Figure 7 display the estimated means $\hat{\mu}_{kr}$ and spatial signal-to-noise ratios $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ within each block. It appears that the spatial activity of the genes in $\mathscr{C}_2$ is largely evident within the internal area of the White Matter $\left(\hat{\tau}_{21}/\hat{\xi}_{21} = 3.45\right)$ and progressively decreases approaching Layer 6 $(\hat{\tau}_{28}/\hat{\xi}_{28} = 1.55$ and $\hat{\tau}_{29}/\hat{\xi}_{29} = 0.58)$. These genes show also a moderate spatial expression on the rest of the cortical area $(\hat{\tau}_{2r}/\hat{\xi}_{2r} \in [0.39,\ 0.9]$, for $r = 2, \ldots,\ 6)$. Last, cluster $\mathscr{D}_7$ denotes a restricted group of spots that are present both within and outside the White Matter, with a non-negligible spatial effect $\left(\hat{\tau}_{27}/\hat{\xi}_{27} = 1.60\right)$. On the contrary, the genes in $\mathscr{C}_1(n_1 = 371)$ show a small spatial variation in every spot cluster expect in $\mathscr{D}_1$ $(\hat{\tau}_{11}/\hat{\xi}_{11} = 0.71$ and $\hat{\tau}_{1r}/\hat{\xi}_{1r} \leq 0.31$ for all $r \neq 1)$, suggesting a constant variation of these genes throughout the cortical area. In fact, $\mathscr{C}_1$ is enriched for housekeeping genes with respect to $\mathscr{C}_2$ (chi-square test, $p = 2.6 \quad \times 10^{-4}$). Housekeeping genes are maintainers of the cellular functions and their activity is not restricted to a specific cell type (Eisenberg and Levanon, 2003). It is therefore expected that these genes show a small spatial variation across the tissue. We notice also from Figure 7 (c) that the estimated means $\{\hat{\mu}_{1r},\ r = 1,\dots,9\}$ are complementary to $\{\hat{\mu}_{2r},\ r = 1,\dots,9\}$: the expression level is smaller within the White Matter area than outside. To ensure that the co-clustering was not driven only by the mean effects, we run also sparseBC using the same number of blocks and $\lambda = 10$: the CER between the gene clusters returned by SpaRTaCo and sparseBC is 0.44, confirming that the two methods perform a substantially different grouping of the data. A further confirmation of the evidence of our gene clustering is given by the very small uncertainty displayed in the left panel of Supplementary Figure 11 (b).

The results discussed above allow us to answer the first two research questions listed in Section 1.1 that motivated our work. We now turn our attention to the third research question, namely the identification of genes that exhibit high specific variation. To do so, for every spot cluster r, we investigate the conditional random variables $\sigma^2_{\widehat{\mathscr{Z}}_i r, i}|\mathrm{x}_{i.}^{\widehat{\mathscr{Z}}_i r}$, for $i = 1, \dots, n$, to determine which genes are most highly variable in each block. We display their density in Supplementary Figures 12 , highlighting in red the twenty genes with the largest $\mathbb{E}\left(\sigma^2_{\widehat{\mathscr{Z}}_i r, i}|\mathrm{x}_{i.}^{\widehat{\mathscr{Z}}_i r}\right)$, for every r. We expect that genes with a large gene-specific variance in some areas are likely to be informative of the biological mechanisms occurring there.

First, we notice that all the most variable genes in each of the nine spot clusters belong to $\mathscr{C}_2$. Among the highly variable genes in $\mathscr{D}_1$, $\mathscr{D}_8$ and $\mathscr{D}_9$ there are *MBP* and *PLP1*, which are responsible, respectively, for the production and the maintenance of myelin, the covering sheath of the nerve fibers in the White Matter. Conversely, among the highly variable genes in $\mathscr{D}_2$ and $\mathscr{D}_7$, we notice *PCP4* and *CCK*: these are markers of distinct subtypes of excitatory neurons present in Layers 5–6 (Hodge et al., 2019). We display the expression of the four genes discussed here in Supplementary Figure 13, showing their pattern in the spot clusters where they appear to be highly variable.

Supplementary Figure 12 highlights some important differences between ranking genes according to the posterior distribution of our gene-specific variance σ_i^2 and the method of Townes et al. (2019) that only ranks genes based on variability without considering the spatial context. This analysis may be used to highlight important genes that would have been missed if the spatial structure of the data would not have been taken into account. Two examples are *CERCAM* and *SAA1*: their ranks according to Townes et al.'s method were 465 and 271, while SpaRTaCo places them among the most variable genes in the White Matter area (cluster $\mathscr{D}_1$) and in a region covering the Layers 3, 5 and 6 (cluster $\mathscr{D}_6$), respectively. We display their expression over the whole tissue in Supplementary Figure 14. *CERCAM* encodes a cell adhesion protein involved in leukocyte transmigration across the blood-brain barrier (Starzyk et al., 2000), while SAA1 is highly expressed in response to inflammation in mouse glial cells (Barbierato et al., 2017).

Taken together, these results convincingly show that our model is able to partition the tissue in coherent clusters, which exhibit cluster-specific gene expression, both spatially coordinated and otherwise, and to detect highly variable genes of potential biological interest in specific areas of the tissue that would not have been found without considering their spatial variability.

6. Discussion.

The growing demand of appropriate statistical methods to analyze spatial transcriptomic experiments has driven us to develop SpaRTaCo, a model-based co-clustering tool that groups genes with a similar profile of spatial expression in specific areas of a tissue. SpaRTaCo brings the concepts of spatial modelling into the co-clustering framework, and thus it can be applied to any dataset with entries in the real domain and whose row or column vectors are multivariate observations recorded at some fixed sites in space. The inference is carried out via maximization of the classification log-likelihood function. To do so, we put together two variants of the EM algorithm, the classification EM and the stochastic EM, forming what we called the classification-stochastic EM. We completed our proposal deriving the formulation of the ICL for our model to drive the model selection.

A series of simulation studies have highlighted that, in the presence of spatial covariance patterns, the major co-clustering models become inadequate to recover the hidden block structure of the data. On the contrary, SpaRTaCo has shown remarkable results in each simulation, managing to distinguish different spatial expression profiles in different areas of the image. It further revealed to be robust to the presence of a nuisance component into the data. The model selection driven by the ICL revealed to be precise but computationally expensive, due to the large number of times the model must be run. On the contrary, other criteria that do not exploit the spatial information of the data are computationally attractive but less accurate. We conclude that the two approaches can be used jointly, using the results given by a fast model selection criterion, such as the PCA-k-means method discussed in Section 4.3, to restrict the range of K and R values to be tested with SpaRTaCo's ICL criterion. Lastly, we demonstrated how our proposal is capable of answering specific biological research questions using a human brain tissue sample processed with the Visium protocol. Our model has identified two clusters of genes with different spatial expression profiles in nine different areas of the tissue. A subsequent downstream analysis has allowed us to determine the highly variable genes in each of the nine pinpointed areas. We additionally showed that some of the genes considered as poorly informative by the deviance method of Townes et al. (2019) are revealed by SpaRTaCo to be highly variable in specific areas of the tissue sample.

Although this article has introduced a complete solution to answer some relevant questions in the analysis of spatial transcriptomics, we believe that there is space for further extensions. To use SpaRTaCo on spatial transcriptomic experiments, the UMI counts must be transformed through a real-valuated function as discussed at the beginning of Section

2. We performed this step using the pre-processing techniques of Townes et al. (2019), which in our application have led to approximately symmetric distributions of the gene expression vectors x_i. In addition, our model is theoretically robust with respect to the presence of heavy tail distributions thanks to the random parameters $\sigma^2_{kr,i}$, that allow to go beyond the normal assumption. Nevertheless, SpaRTaCo could be extended to directly model UMI counts, similarly to how SPARK (Sun, Zhu and Zhou, 2020) has extended SpatialDE (Svensson, Teichmann and Stegle, 2018). Second, to overcome the limitations of the stochastic EM presented in Section 4.5, we could explore the *simulated annealing* algorithm (Van Laarhoven and Aarts, 1987), to reduce the chances of converging to local maxima.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

The authors are thankful to the Editor, the Associate Editor, and the two Reviewers for their careful evaluation of our work and for their precious comments, to Giovanna Menardi and Alessandro Casa for the precious discussions on co-clustering and to Levi Waldron and Vince Carey for help with the framing of the biological questions. We finally thank Dario Righelli for his help with the software implementation.

Funding.

This work was supported in part by CZF2019–002443 (DR) from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. The authors are supported by the National Cancer Institute of the National Institutes of Health (U24CA180996).

REFERENCES

ALLEN GI and TIBSHIRANI R (2010). Transposable regularized covariance models with an application to missing data imputation. The Annals of Applied Statistics 4 764 – 790. [PubMed: 26877823]

ANDERLUCCI L and VIROLI C (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. The Annals of Applied Statistics 9 777–800.

BARBIERATO M, BORRI M, FACCI L, ZUSSO M, SKAPER SD and GIUSTI P (2017). Expression and differential responsiveness of central nervous system glial cell populations to the acute phase protein serum amyloid A. Scientific reports 7 1–14. [PubMed: 28127051]

BENJAMINI Y and HOCHBERG Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 57 289–300.

BIERNACKI C, CELEUX G and GOVAERT G (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence 22 719–725.

BOUVEYRON C, BOZZI L, JACQUES J and JOLLOIS F-X (2018). The functional latent block model for the co-clustering of electricity consumption curves. J. R. Stat. Soc. Ser. C. Appl. Stat 67 897–915.

BOUVEYRON C, CELEUX G, MURPHY TB and RAFTERY AE (2019). Model-based clustering and classification for data science. Cambridge Series in Statistical and Probabilistic Mathematics Cambridge University Press, Cambridge With applications in R.

BYRD RH, LU P, NOCEDAL J and ZHU CY (1995). A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16 1190–1208.

CAPONERA A, DENTI F, RIGON T, SOTTOSANTI A and GELFAND A (2017). Hierarchical Spatio-Temporal Modeling of Resting State fMRI Data. In START UP RESEARCH 111–130. Springer.

CASA A, BOUVEYRON C, EROSHEVA E and MENARDI G (2021). Co-clustering of Time-Dependent Data via the Shape Invariant Model. Journal of Classification

CELEUX G and GOVAERT G (1992). A classification EM algorithm for clustering and two stochastic versions. Computational statistics & Data analysis 14 315–332.

CHEN KH, BOETTIGER AN, MOFFITT JR, WANG S and ZHUANG X (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348 aaa6090. [PubMed: 25858977]

CRESSIE N (2015). Statistics for spatial data John Wiley & Sons.

DE LA CRUZ-MESÍA R and MARSHALL G (2006). Non-linear random effects models with continuous time autoregressive errors: a Bayesian approach. Statistics in medicine 25 1471–1484. [PubMed: 16013034]

DELATTRE M, LAVIELLE M, POURSAT M-A et al. (2014). A note on BIC in mixed-effects models. Electronic journal of statistics 8 456–475.

DRIES R, ZHU Q, DONG R, ENG C-HL, LI H, LIU K et al. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome biology 22 1–31. [PubMed: 33397451]

EDSGÄRD D, JOHNSSON P and SANDBERG R (2018). Identification of spatial expression trends in single-cell gene expression data. Nature methods 15 339–342. [PubMed: 29553578]

EFRON B (2009). Are a Set of Microarrays Independent of Each Other? The Annals of Applied Statistics 3 922–942. [PubMed: 20563291]

EISENBERG E and LEVANON EY (2003). Human housekeeping genes are compact. TRENDS in Genetics 19 362–365. [PubMed: 12850439]

GOVAERT G and NADIF M (2008). Block clustering with Bernoulli mixture models: comparison of different approaches. Computational Statistics & Data Analysis 52 3233–3245.

GOVAERT G and NADIF M (2010). Latent block model for contingency table. Communications in Statistics. Theory and Methods 39 416–425.

GOVAERT G and NADIF M (2013). Co-clustering: models, algorithms and applications John Wiley & Sons.

GUPTA AK and NAGAR DK (2018). Matrix variate distributions 104. CRC Press.

HODGE RD, BAKKEN TE, MILLER JA, SMITH KA, BARKAN ER, GRAYBUCK LT et al. . (2019). Conserved cell types with divergent features in human versus mouse cortex. Nature 573 61–68. [PubMed: 31435019]

KERIBIN C, BRAULT V, CELEUX G and GOVAERT G (2015). Estimation and selection for the latent block model on categorical data. Statistics and Computing 25 1201–1216.

LUBECK E, COSKUN AF, ZHIYENTAYEV T, AHMAD M and CAI L (2014). Single-cell in situ RNA profiling by sequential hybridization. Nature methods 11 360–361. [PubMed: 24681720]

MARX V (2021). Method of the Year 2020: spatially resolved transcriptomics. Nature Methods 18 9–14. [PubMed: 33408395]

MAYNARD KR, COLLADO-TORRES L, WEBER LM, UYTINGCO C, BARRY BK, WILLIAMS SR et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature Neuroscience

MORAN GE, ROČ KOVÁ V and GEORGE EI (2021). Spike-and-slab Lasso biclustering. The Annals of Applied Statistics 15 148 – 173.

MURUA A and QUINTANA FA (2021). Biclustering via Semiparametric Bayesian Inference. Bayesian Analysis 1 – 27.

NOBILE A and FEARNSIDE AT (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. Statistics and Computing 17 147–162.

PARDO B, SPANGLER A, WEBER LM, HICKS SC, JAFFE AE, MARTINOWICH K et al. (2021). spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. bioRxiv

RAO N, CLARK S and HABERN O (2020). Bridging Genomics and Tissue Pathology. Genetic Engineering & Biotechnology News 40 50–51.

RASMUSSEN CE and WILLIAMS CKI (2006). Gaussian Processes for Machine Learning The MIT Press.

RIGHELLI D, WEBER LM, CROWELL HL, PARDO B, COLLADO-TORRES L, GHAZANFAR S et al. (2021). SpatialExperiment: infrastructure for spatially resolved transcriptomics data in R using Bioconductor. bioRxiv

RODRIQUES SG, STICKELS RR, GOEVA A, MARTIN CA, MURRAY E, VANDERBURG CR et al. . (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 363 1463–1467. [PubMed: 30923225]

SMYTH GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3.

SOTTOSANTI A and RISSO D (2022). Supplementary to "Co-clustering of Spatially Resolved Transcriptomic Data"

STARZYK RM, ROSENOW C, FRYE J, LEISMANN M, RODZINSKI E, PUTNEY S and TUOMANEN EI (2000). Cerebral cell adhesion molecule: a novel leukocyte adhesion determinant on blood-brain barrier capillary endothelium. The Journal of Infectious Diseases 181 181–187. [PubMed: 10608765]

SUN S, ZHU J and ZHOU X (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature methods 17 193–200. [PubMed: 31988518]

SVENSSON V, TEICHMANN SA and STEGLE O (2018). SpatialDE: identification of spatially variable genes. Nature methods 15 343–346. [PubMed: 29553579]

TAN KM and WITTEN DM (2014). Sparse biclustering of transposable data. Journal of Computational and Graphical Statistics 23 985–1008. [PubMed: 25364221]

TOWNES FW, HICKS SC, ARYEE MJ and IRIZARRY RA (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome biology 20 1–16. [PubMed: 30606230]

VAN LAARHOVEN PJ and AARTS EH (1987). Simulated annealing. In Simulated annealing: Theory and applications 7–15. Springer.

WITTEN DM and TIBSHIRANI R (2009). Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society: Series B (Methodological) 71 615–636.

WITTEN DM and TIBSHIRANI R (2010). A Framework for Feature Selection in Clustering. Journal of the American Statistical Association 105 713–726. [PubMed: 20811510]

WYSE J and FRIEL N (2012). Block clustering with collapsed latent block models. Statistics and Computing 22 415–428.

ZHAO E, STONE MR, REN X, GUENTHOER J, SMYTHE KS, PULLIAM T et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. Nature Biotechnology 1–10.

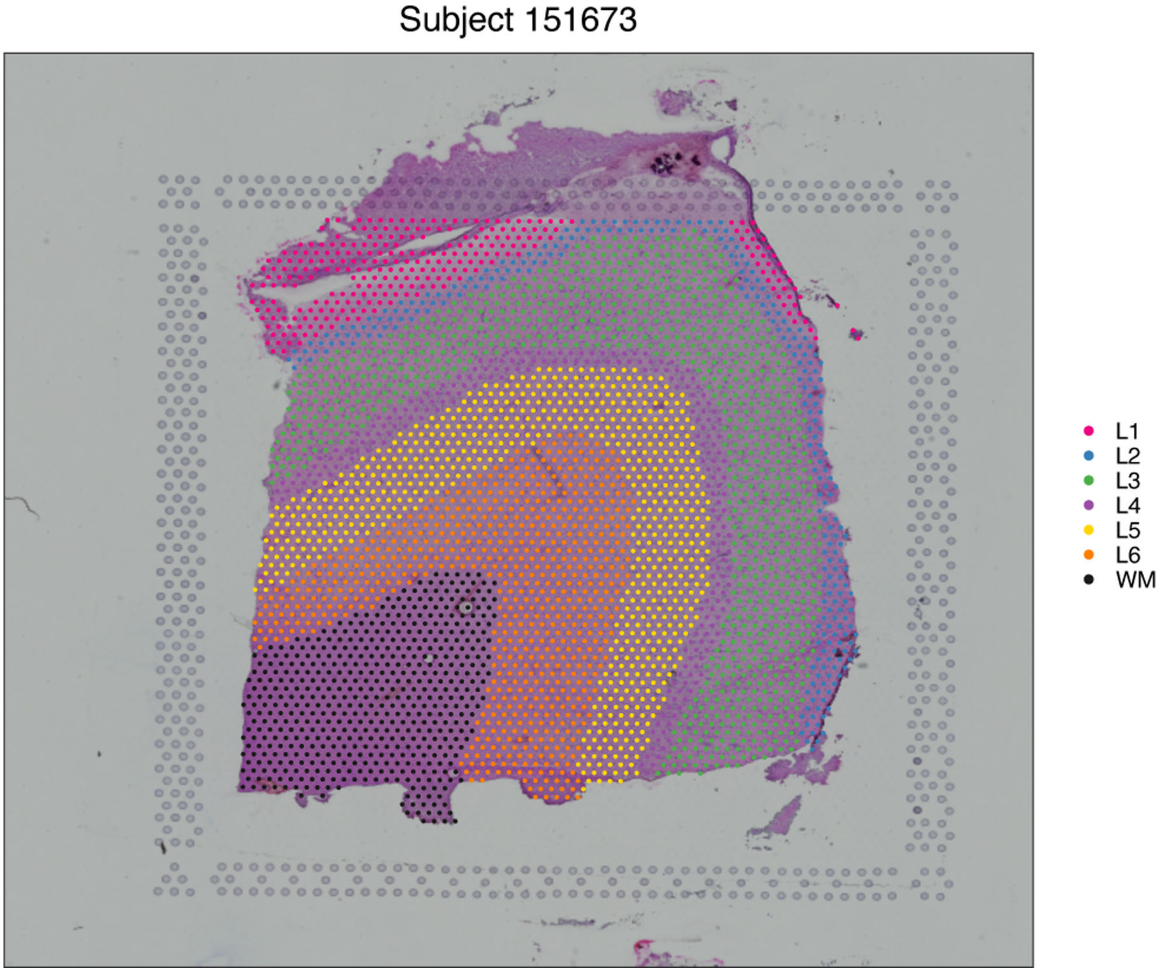

FIG 1.

Tissue sample of LIBD human dorsolateral prefrontal cortex (DLPFC) processed with Visium platform and stored in the R package `spatialLIBD`. The dots represent the spots over the chip surface. Different colors denote a manual annotation of the areas performed by Maynard et al. (2021): they recognize a White Matter (WM) stratum in the bottom-left part of the image, and 6 Layers (from L6 to L1) moving toward the top-right.

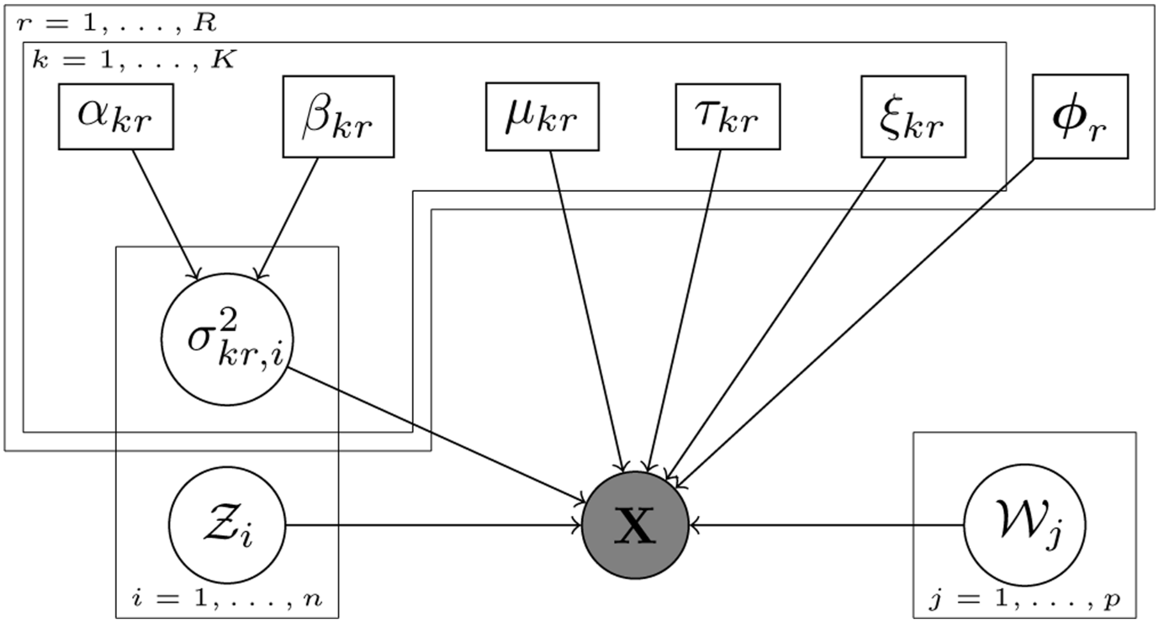

FIG 2.

DAG of the SpaRTaCo co-clustering model. Grey circle denotes the data, white circles are the latent random variables, and white rectangles are the model parameters.

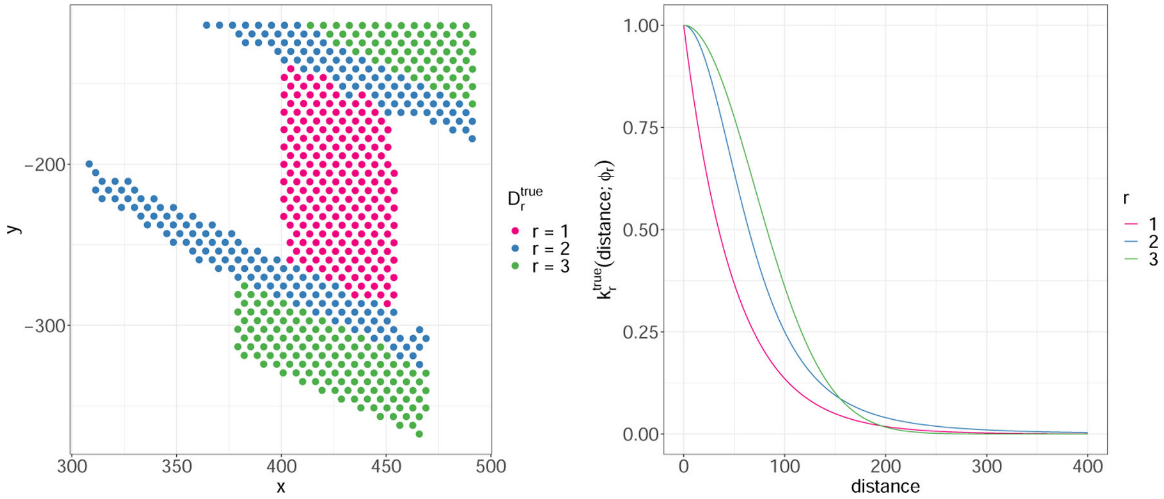

FIG 3.

Left: map of the spots used to generate the simulation experiments, extracted from the subject 151507 contained in the package `spatialLIBD`. The clusters are of equal size, $p_1 = p_2 = p_3 = 200$. Right: comparison of the covariance functions used in the three clusters of spots. When $r = 1$, the covariance is Exponential with scale $\theta_E = 50$, when $r = 2$, it is Rational Quadratic with $\theta_R = 50$ and $\alpha_R = 2$, and when $r = 3$ it is Gaussian with scale $\theta_G = 70$.

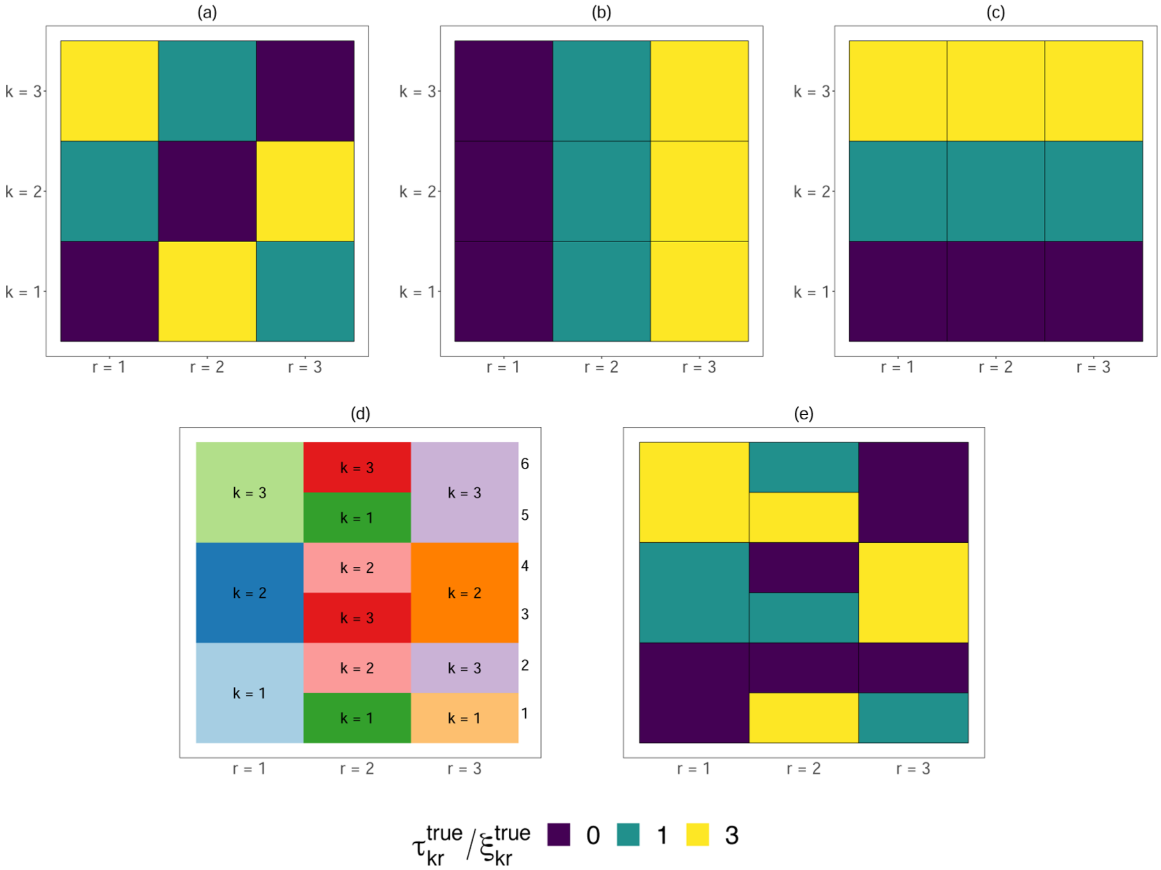

FIG 4.

Representation of the latent block structures used to generate the simulation experiments. All the blocks in Panels (a)-(c) have the same size and are colored according to the value of the spatial signal-to-noise ratio $\tau_{kr}^{\text{true}}/\xi_{kr}^{\text{true}}$. The setup in Panel (a) is used in Sections 4.3 and 4.6, Panel (b) is used in Section 4.4, Panel (c) in Section 4.5 and Panel (e) in Section 4.7. Panel (d) gives the hidden block structure of Simulation 4.7. Within the columns 1 and 2, the row clusters have the same size (200), while in the third column it is $n_{13} = 100$, $n_{23} = 200$ and $n_{33} = 300$. The numbers from 1 to 6 on the right denote the alternative clusters $\mathscr{C}_1^{*\text{true}},\ldots,\mathscr{C}_6^{*\text{true}}$.

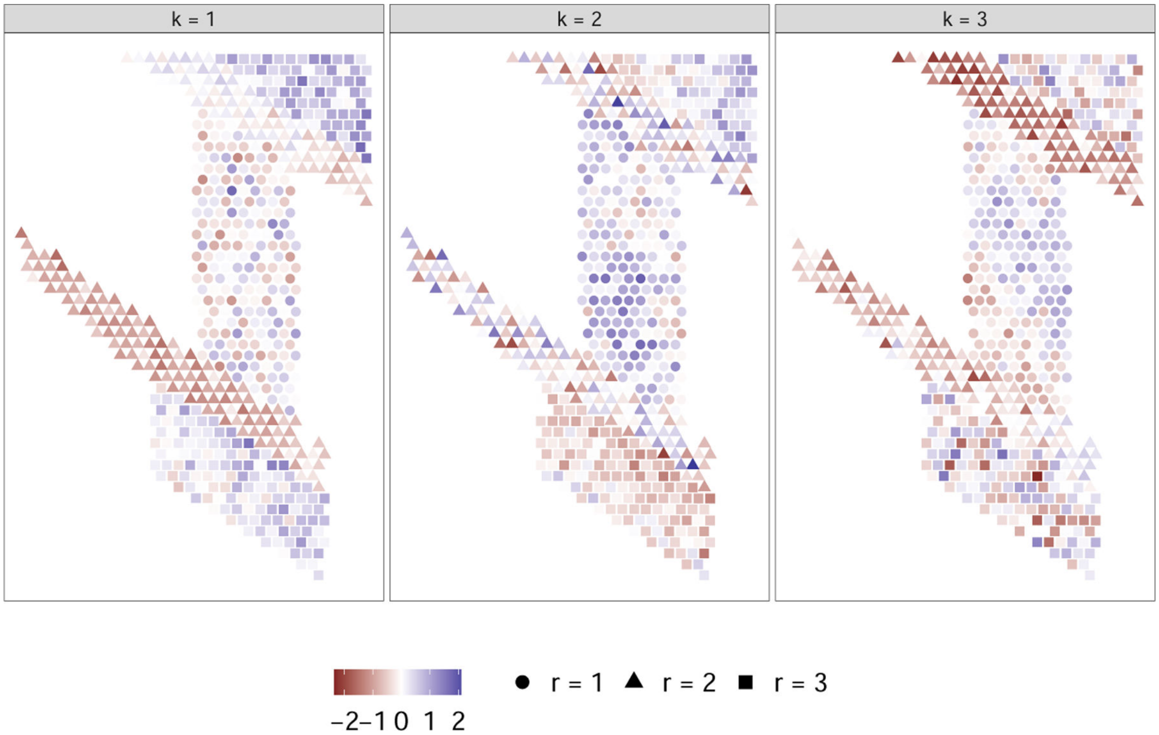

FIG 5.

Examples of a spatial experiment generated under Simulation 1. The spots are coloured according to $n_k^{-1}\left(\mathrm{X}^{k\cdot}\right)^T 1_{n_k}$, the average expression of the k-th gene cluster. The three spot clusters are displayed with different symbols. The co-clusters with no spatial expression are $(k = 1, r = 1)$, $(k = 2, r = 2)$ and $(k = 3, r = 3)$, and the co-clusters with the largest spatial signal-to-noise ratio are $(k = 1, r = 2)$, $(k = 2, r = 3)$ and $(k = 3, r = 1)$.

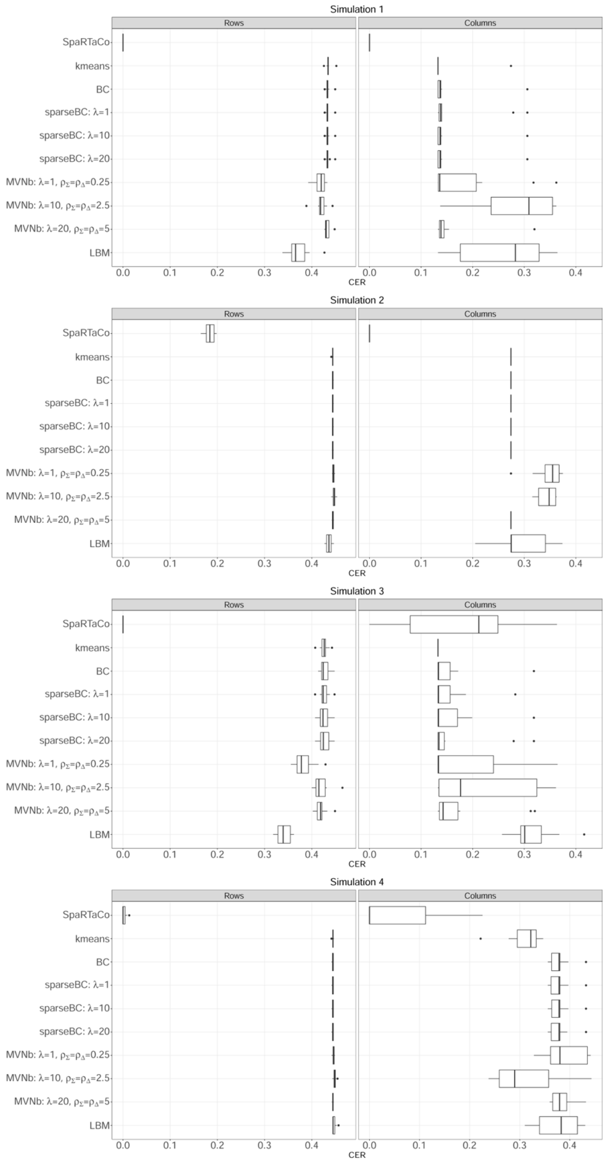

FIG 6.

Results from Simulations 1–4. For each scenario, we generated 10 datasets and we applied the co-clustering models listed in Section 4.2. Every figure gives the boxplots of the CER obtained on the rows and on the columns.

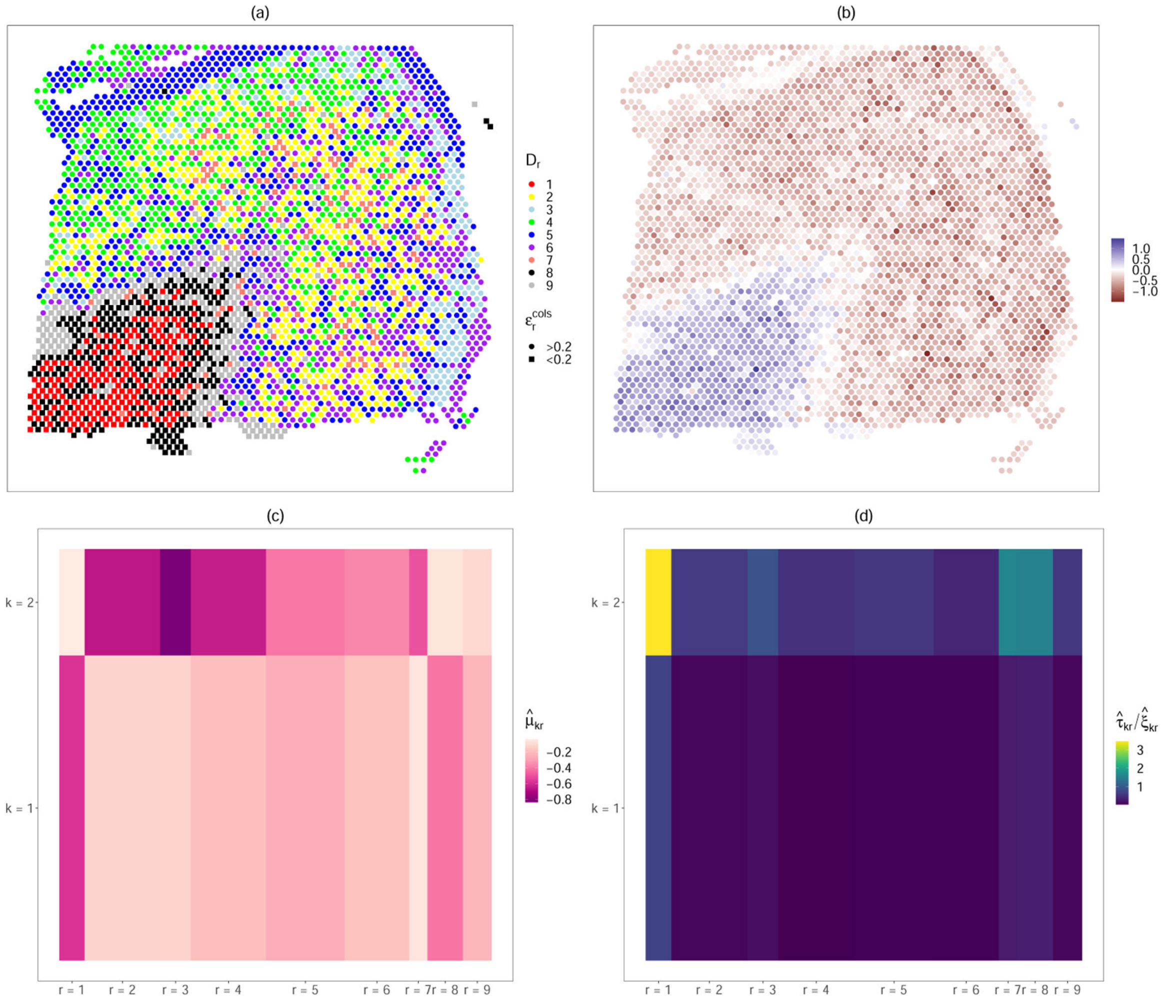

FIG 7.

Results on the human dorsolateral prefrontal cortex data. The first row displays the 3,639 spots: in Panel (a) they are colored according to the clusters returned by SpaRTaCo and shaped according to the clustering uncertainty $\varepsilon_r^{\text{cols}}$, in Panel (b) they are colored according to the average gene expression in the estimated cluster $\mathscr{C}_2$. Panels (c) and (d) represent the data matrix tessellated into the 18 discovered blocks. Both the genes and the spots are reordered based on the estimated clusters for visualization purposes. The graphs are colored according to the estimated mean $\hat{\mu}_{kr}$ (c) and to the estimated spatial signal-to-noise ratio $\hat{\tau}_{kr}/\hat{\xi}_{kr}$ (d).